education sciences

MDPI

Review

Teacher Mental Health and Well-Being from the Pandemic to the Post-Pandemic Era: A Systematic Review of Interventions, Policies, and Persistent Gaps

Zhili Wang [1,2] and Baoxin Long [1,3,*]

[1] Faculty of Education, Shaanxi Normal University, Xi'an 710062, China; 120200108@taru.edu.cn
[2] College of Humanities, Tarim University, Alar 843300, China
[3] School of Minzu Education, Shaanxi Normal University, Xi'an 710062, China
* Correspondence: longbx@snnu.edu.cn

Abstract

Teachers' mental health problems have remained persistently elevated since the pandemic, yet system-level responses remain structurally inadequate. This systematic review synthesizes 66 peer-reviewed studies and policy reports (2020–2026). Meta-analyses of 54 studies (256,896 teachers, 22 countries) found stress prevalence at 62.6%, anxiety at 36.3%, and depression at 59.9%. Longitudinal evidence shows no recovery across 2021–2026; teachers fare worse than comparable professionals on all well-being indicators. Interventions show moderate effects (g = 0.42; trim-and-fill-adjusted g = 0.34), but the latter is an exploratory sensitivity analysis suggesting small-study effects, not a precise publication-bias estimate. Out of 24 intervention studies, 91.7% target individual-level coping; no peer-reviewed quantitative study evaluates macro-structural reforms. We propose the Evidence Hierarchy Mismatch Theory (EHMT) as one plausible explanation, alongside competing explanations. We introduce a three-tier framework (macro–meso–micro) extending the job demands–resources model with structural–social justice considerations. Moral injury is proposed as a speculative candidate mechanism linking structural constraints to burnout and attrition, not as a replacement for established constructs. The military–education analogy is heuristic, not ontological. Normatively, we call for a paradigm shift from individual remediation to systemic reform, methodological pluralism, mandatory null-results publication, and sustainability as an evaluation criterion. Empirical findings, theoretical interpretation, and normative proposals are labeled separately.

Keywords: teacher mental health; post-pandemic; social justice; moral injury; evidence hierarchy; well-being

Check for updates

Academic Editor: Pearl Subban

Received: 5 August 2026
Revised: 28 September 2026
Accepted: 29 September 2026
Published: 3 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Teaching is inherently stressful, and over the past five years, workload, boundary blurring, and emotional labor have intensified to unsustainable levels. Longitudinal evidence confirms that teacher distress persists and exceeds that of comparable professionals on every well-being indicator (Pressley, 2021; Steiner et al., 2025). Consequences extend to student achievement, motivation, and educational equity (Kraft & Lyon, 2022). Gaps remain in understanding (a) intervention effectiveness across ecological levels, (b) policy adequacy, and (c) structural barriers to sustainable change. This review synthesizes evidence across these domains and critiques the evidence hierarchy that may bias resources toward low-impact individual interventions. We investigate whether persistence of teacher distress

is better explained by how the field defines, studies, and responds to the problem than by individual resilience deficits. This is a guiding hypothesis, not a predetermined conclusion.

For this review, the pandemic period is 2020–2022 and the post-pandemic recovery period is 2023–2026. Studies from both periods were eligible because the review examines persistence and recovery; primary synthesis and policy analysis focus on 2023–2026, with sensitivity analyses in Section 3.1 and Supplementary Table S1. This manuscript has two functions: (a) a systematic review of empirical evidence on teacher mental health, and (b) development of a theoretical framework (Three-Tier Integrative Framework and EHMT). To avoid conflating these, we use the following explicit labels: [empirical finding], [theoretical interpretation], [policy implication], [normative proposal], and [speculative hypothesis]. Section 2 describes the method; Section 3 reports empirical findings only; Section 4 develops theoretical interpretation, policy implications, and normative proposals. Section 4.5 grades each proposition, and Supplementary Table S2 maps each component based on its method, evidence standard, and interpretive status.

1.1. Theoretical Framework: A Three-Tier Integrative Model

This section is conceptual and develops a testable framework, not an established causal model. Existing models are too narrow (JD-R explains individual-level job characteristics but lacks structural and moral dimensions), too abstract (Bronfenbrenner's ecology describes contexts but offers no mechanism for resource allocation), or too critical without prescription (social justice scholarship diagnoses inequity but rarely specifies intervention loci). We integrate these as nested, sequentially ordered strata. JD-R specifies proximal psychological mechanisms (demands → strain → burnout) (Bakker & Demerouti, 2017); Bronfenbrenner maps multi-level contexts; and a structural–social justice lens identifies inequities by race, class, and gender that JD-R's value-neutral formulation obscures. The result is explanatory, ecological, and critical.

Three tiers are distinguished by primary locus of intervention, not by the nature of the problem:

Macro-level (structural): This includes policy regimes, funding inequities, and systemic discrimination. These allocate demands and resources unequally; per-pupil expenditure predicts administrative support staffing ratios (Mccarthy et al., 2025). [Theoretical construct].

Meso-level (organizational): This includes school leadership, administrative support, professional culture, and working conditions. These are shaped by macro-level decisions and mediate structural conditions in daily experience. [Theoretical construct].

Micro-level (individual): This includes coping, self-efficacy, emotional regulation, and psychological resources. These are proximal influences but are shaped by macro and meso contexts. [Theoretical construct].

A micro-level intervention targets the individual teacher regardless of problem origin. This precision frames the mismatch documented in Section 3.2: most interventions target the micro level, while drivers may operate primarily at macro and meso levels. The hypothesized downward pathway (macro → meso → micro) proposes that funding cuts reduce administrative support, increasing workload and eroding well-being. This is consistent with observational evidence but not causally confirmed. Upward feedback (micro → macro) is hypothesized to be conditional on union strength, media framing, and political receptivity; where these are weak, the loop may remain fractured—the "upward fracture" (P3). This is an active mechanism of responsibility diffusion, not passive absence: individual distress is psychologized (Zembylas, 2021) and structural grievances are translated into personal deficits, preempting collective policy claims. This occurs within a neoliberal

political economy of accountability, performativity, and eroded autonomy (Acuña, 2024). Figure 1 illustrates the framework.

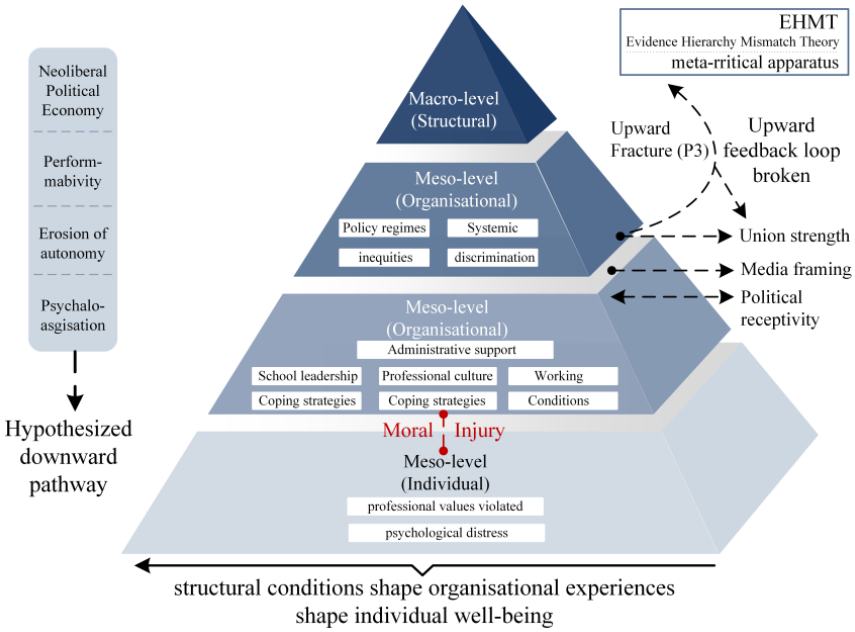

Figure 1. Three-Tier Integrative Framework. Note. Solid arrows: Evidence-based associations in included studies. Dashed arrows: Proposed explanatory pathways, not causally confirmed. Dotted arrows: Speculative mechanisms requiring empirical testing.

1.1.1. Moral Injury as a Candidate Bridging Construct

We propose moral injury as a testable hypothesis, not an established finding. Two claims are distinguished. The first is an observed-evidence claim: qualitative studies document teachers describing betrayal, guilt, shame, and value violation when institutional constraints prevent fulfillment of professional duties of care (Sokal et al., 2020; Loh & Hu, 2014). The second is a hypothesized mediation claim: moral injury mediates structural constraints and burnout/attrition. The first is empirically anchored but limited to qualitative and cross-sectional accounts; the second remains speculative and untested with validated instruments or longitudinal mediation.

Teacher moral injury is defined as psychological distress (guilt, shame, betrayal, and demoralization) when teachers are compelled by policies, directives, or resource constraints to violate professional values and duty of care. It is incremental to burnout, demoralization, work stress, emotional exhaustion, and compassion fatigue. Moral injury is incremental to burnout, demoralization, work stress, emotional exhaustion, and compassion fatigue; a full construct comparison is provided in Supplementary Table S3. Operationalization: the Moral Injury Events Scale (MIES) was adapted as T-MIES, with dimensions including betrayal by institution/leadership/policy, guilt over constrained care, shame regarding professional identity, moral distress from value violation, and demoralization from repeated value violation. Validation requires content validity with teachers and teacher educators; discriminant validity against burnout (MBI), depression (PHQ-9), anxiety (GAD-7), and compassion fatigue (ProQOL); incremental validity predicting burnout/attrition; and three-wave longitudinal mediation. Current evidence remains qualitative and cross-sectional; no validated teacher-specific instrument or longitudinal mediation exists. Moral injury is therefore graded speculative in Section 4.5.

1.1.2. Limits of the Military–Education Analogy

The analogy is heuristic, not ontological. We do not equate teaching with combat or claim identical moral injury. Four limits apply. The first limit is purpose and role: military moral injury involves existential threat, lethal force, and command responsibility; teaching involves developmental care, pedagogy, and professional judgment. The second limit is organizational structure: militaries are command hierarchies; schools are professional bureaucracies with distributed autonomy. Betrayal in teaching may appear as policy contradiction, resource scarcity, administrative indifference, or accountability pressure. The third limit is exposure: military moral injury often involves discrete high-intensity events; teacher moral injury is chronic, cumulative, and institutional. The fourth limit involves consequences and measurement: military moral injury is linked to PTSD, depression, and suicide risk; teacher distress is expressed as burnout, demoralization, guilt, shame, and attrition. Instruments cannot transfer without adaptation. School governance may condition expression—high-stakes accountability may produce metric-driven value violation, while low-stakes systems may differ—but comparative evidence is hypothesis-generating only. Conceptual overreach risks pathologizing teaching, militarizing discourse, obscuring moral rewards, conflating constructs, and importing unadapted measures. Moral injury remains a bounded, provisional candidate mechanism.

1.2. Distinction from Prior Reviews and Evidence Hierarchy Mismatch Theory

Table 1 summarizes key distinctions from prior reviews.

Table 1. Comparison with prior systematic reviews.

Prior Review	Scope	This Review's Distinct Contribution
Aziku and Zhang (2024)	Pandemic publication trends up to 2023; no intervention or policy evaluation	2021–2026 coverage + prevalence + interventions + policy + meta-theoretical critique
Berger et al. (2022)	Intervention studies only; no prevalence or policy context	Three-tier Integrated Framework (prevalence + interventions + policy)
Silva et al. (2021); Ozamiz-Etxebarria et al. (2021)	Acute pandemic prevalence; no post-pandemic tracking	Post-pandemic persistence analysis + structural justice lens + EHMT

EHMT is a meta-theoretical proposition. The empirical finding it explains is the tier distribution in Section 3.2: 91.7% of intervention studies target the micro level and 0% evaluate macro-level reforms. EHMT itself is not empirical. It comprises three claims:

C1 Descriptive [empirical]: Published intervention evidence is systematically skewed toward micro-level interventions and away from macro-level reforms.

C2 Explanatory [theoretical]: This skew is hypothesized to arise from epistemological, institutional, and practical biases. Epistemological bias is uncritical transplantation of the medical RCT hierarchy; institutional bias is funding and publication incentives favoring short-term, controllable studies; practical bias is RCT infeasibility for system-level reforms.

C3 Consequential [theoretical]: This may lead to epistemic foreclosure: macro-reforms remain under-evidenced, under-funded, and rarely evaluated.

Only C1 is supported by the tier distribution; C2–C3 are theoretical. The critique is not of RCTs as a method but of uncritical hierarchy treating RCTs as universally superior (Coppe et al., 2025; Parra & Edwards, 2024). RCTs remain valuable where randomization is feasible, ethical, and appropriate. The paradox is that macro-structural reforms (salary equalization, class-size mandates, and accountability redesign) may have the greatest population-level potential but are RCT-infeasible, while micro-interventions (mindfulness and CBT) are

RCT-easy but may have minimal sustainable systemic impact. An "inverse evidence law" may operate (Nutbeam, 2004): we may know the least about interventions with the greatest population potential. Quasi-experimental designs—interrupted time series, difference-in-differences, and comparative case studies—are appropriate and complementary (Petticrew & Roberts, 2003).

Five competing explanations must be considered: financial limitations; measurement difficulty; publication patterns; policy data access; and political difficulties. EHMT is therefore one plausible explanation, and is not proven. Its distinctive prediction—that even when controlling these factors, the evidence hierarchy independently reduces prioritization of macro-level quasi-experimental evaluation—remains untested. Table S4 summarizes these explanations. EHMT is bounded to evaluative evidence cultures in high-income, Anglophone-dominated systems; it does not explain policy adoption or implementation fidelity.

1.3. Research Questions and Propositions

Propositions are hypotheses to be evaluated, not conclusions established in advance. Their empirical weight is assessed in Section 4.5.

RQ1: What is the prevalence and nature of teacher mental health difficulties in the post-pandemic era?

P1 (Hypothesized Downward Pathway). Macro-level funding inequities predict meso-level resource constraints, which predict micro-level distress, mediated by workload and accountability pressure. Individual links are supported by cross-sectional and multi-level evidence, but the complete macro–meso–micro mediation chain has not been tested longitudinally or causally. Grade: Moderate.

RQ2: What interventions have been implemented, and what is the evidence for their effectiveness, with attention to tier distribution and sustainability?

P4 (Evidence Hierarchy Mismatch). The distribution of published intervention studies is systematically biased toward micro-level interventions due to epistemological, institutional, and practical biases, not population-level impact assessments. See C1–C3. P4a (descriptive): Moderate. P4b (explanatory): Speculative.

P5 (Macro-Level Effectiveness). Macro-level structural reforms may produce greater population-level reductions in teacher distress than micro-level interventions. No peer-reviewed study in the identified quantitative effectiveness literature has evaluated a macro-level reform; comparative effectiveness remains untested. Grade: Speculative.

RQ3: What policy responses have been developed, and what persistent gaps remain?

P3 (Hypothesized Upward Fracture). The upward feedback loop (micro → macro) is moderated by (P3a) union density, (P3b) media framing, and (P3c) political receptivity. Where these are weak, the loop remains fractured. This is an active mechanism of systemic responsibility diffusion, not passive absence, and is institutionally produced and discursively maintained. Grade: Weak.

P2 (Moral Injury as Candidate Mediator). Resource constraints and accountability demands predict teacher moral injury, which mediates the relationship between working conditions and burnout/attrition. Grade: Speculative. Its theoretical function is to explain why individual-level interventions may systematically fail, thereby reinforcing EHMT.

The propositions vary in empirical weight, from moderate (P1, P4a) to weak (P3) and speculative (P2, P4b, P5). Section 4.5 provides explicit grading to distinguish established findings from directional inferences and untested hypotheses.

2. Method

2.1. Search Strategy and Eligibility

This systematic review followed PRISMA 2020. We searched Web of Science, PsycINFO, PubMed, Embase, ERIC, and Scopus for peer-reviewed articles, systematic reviews, and policy reports published in January 2021–June 2026. The search yielded 2081 records; after duplicate removal, 1847 remained (Supplementary Table S5). Pandemic period = 2020–2022; post-pandemic recovery = 2023–2026. Because the search began in January 2021, 2020 studies were eligible only if they were indexed during the search window; this limitation is noted in Section 4.6.

Inclusion criteria are as follows: (a) including teachers, educators, or school staff in primary, secondary, or higher education; (b) exploring mental health, well-being, stress, burnout, anxiety, or depression; (c) set during the pandemic or post-pandemic period; (d) peer-reviewed journals or official policy reports; (e) written in English; (f) containing original empirical findings, systematic reviews, or policy analyses. Exclusions are as follows: student-only outcomes; opinion pieces/editorials; pre-2020 publications; no pandemic context. Equity-relevant evidence was prioritized where multiple eligible studies addressed the same question; social justice is treated as an analytical framework, not a normative overlay (Section 2.8). Periods were defined a priori; studies spanning both were classified by the primary data-collection window. Where a study reported data from both periods, it contributed to descriptive counts but only once to the meta-analysis, using the post-pandemic wave where available. Mental health, well-being, distress, burnout, anxiety, and depression are treated as distinct constructs; discrepancies are noted in Supplementary Table S6.

2.2. Study Selection and PRISMA Flow

Title/abstract screening and full-text review were conducted independently by two reviewers ($\kappa = 0.84$ screening; $\kappa = 0.79$ full-text). Of the 1847 records, 423 were excluded during title/abstract screening (irrelevant topics, non-teacher populations, and non-empirical studies), leaving 1424 for the full-text assessment. Of the 1424 reports sought, 1287 were not retrieved (unavailable, conference abstracts only, or not accessible in English). The remaining 137 full texts were assessed; 71 were excluded (no teacher-specific outcomes, n = 24; no pandemic context, n = 18; opinion pieces, n = 15; pre-2020, n = 14), yielding 66 studies (Figure 2). The PRISMA 2020 checklist is included as Supplementary Table S7; excluded studies with reasons are included in Supplementary Table S8. Geographic distribution is as follows: US n = 22; Europe n = 18; Australia n = 8; Asia n = 10; international/multi-country n = 8.

From the 66 studies, we identified 24 intervention effectiveness studies after removing pure prevalence (n = 28), correlational/predictor (n = 9), and policy/discourse analyses (n = 5). A total of 24 met additional criteria: an intervention and control group (or pre–post design with $\geq$3 time points); a quantitative teacher mental health/well-being outcome; duration $\geq$1 week; sample $\geq$15 per group. Exclusion decisions were made by two reviewers ($\kappa = 0.81$), with disagreements resolved by consensus. The 24 studies included 8742 teachers (range 28–1247), conducted in the US (41.7%), Europe (25.0%), Asia (20.8%), and Australia/Oceania (12.5%); only 2 (8.3%) reported implementation in low-income or high-poverty districts. Supplementary Table S9 lists these studies with tier and sustainability coding. Supplementary Table S6 categorizes the 66 studies: 28 prevalence, 24 intervention effectiveness, 9 correlational/predictor, and 5 policy/discourse analyses. Systematic reviews and meta-analyses contextualized evidence but were not counted as primary studies.

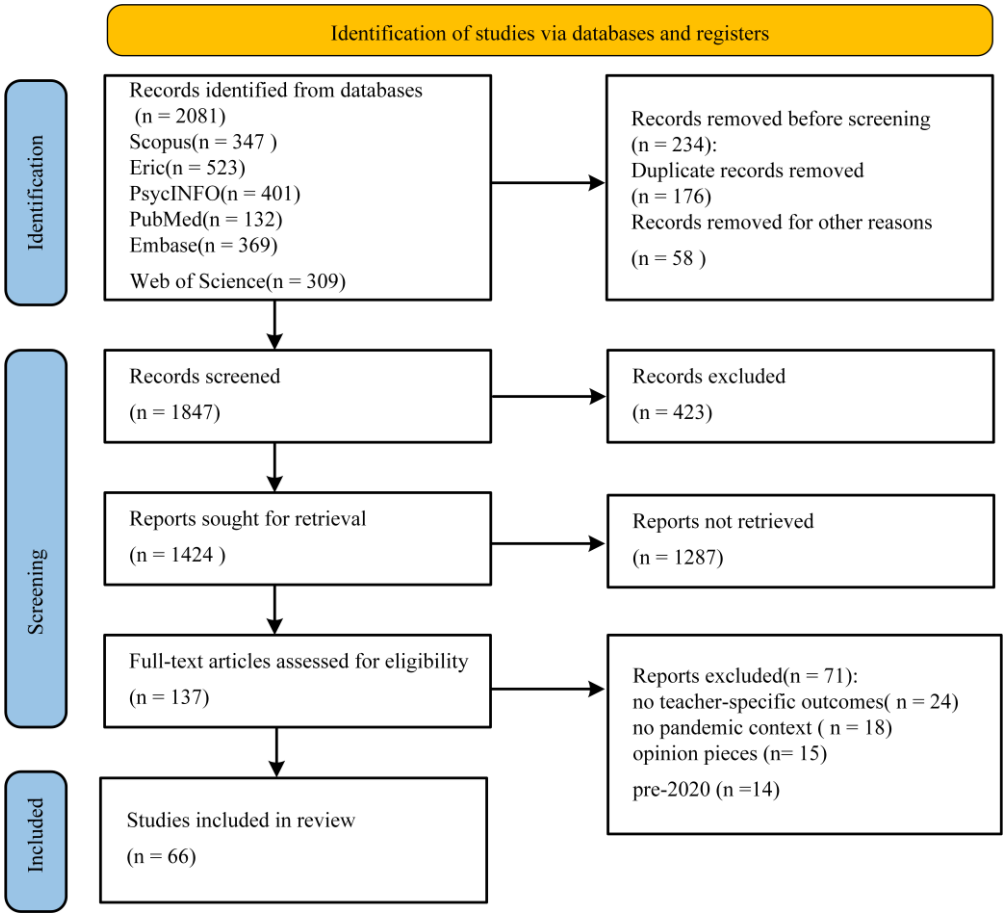

Figure 2. Flow diagram of the literature search according to the PRISMA 2020 guidelines.

2.3. Data Extraction and Synthesis

Data were extracted using a standardized form (Supplementary Table S10). A narrative synthesis (SWiM guidelines) organized studies around the Three-Tier Framework's interactional logic rather than a linear prevalence–interventions–policy sequence. Equity-relevant variables were extracted where reported (school poverty, student–teacher ratio, administrative support staffing, per-pupil expenditure, teacher race/ethnicity, gender, career stage, union density, policy targeting, and subgroup outcomes; Supplementary Table S11) and used in equity-weighted interpretation (Sections 3.4 and 4.4). Because most included studies are cross-sectional, observational, policy-reporting, or narrative-synthesis designs, this review does not identify causal pathways. Causal language is used only for explicitly labeled hypotheses; "predict" is used in a statistical-association sense unless explicitly framed as a causal hypothesis. Normative statements are labeled throughout; wording is calibrated to the grades used in Section 4.5 (Supplementary Table S12).

This review integrates four evidence streams: (1) prevalence evidence (RQ1); (2) intervention evidence from 24 effectiveness studies, including effect-size synthesis, subgroup analyses, publication-bias sensitivity analyses, and tier coding (RQ2); (3) policy evidence from 43 gray literature documents, coded for enforceability, evaluation status, equity targeting, and failure mechanisms (RQ3); and (4) a theoretical synthesis developing the Three-Tier Integrative Framework, EHMT, and propositions P1–P5. Integration occurs at three points: descriptive integration maps findings onto macro–meso–micro tiers, producing the tier distribution (91.7% micro, 8.3% meso, 0% macro) and policy-failure mechanisms in Section 3.3 explanatory integration uses EHMT to interpret tier distribution and policy gaps, with competing explanations in Sections 1.2 and 4.3 and propositional integration

derives P1–P5, graded in Table 4, with moral injury as a speculative bridging hypothesis (P2). Supplementary Table S13 provides the integration matrix; Figure 3 illustrates the workflow. Full data and analysis scripts are given in Supplementary Tables S9, S11, S13 and S14.

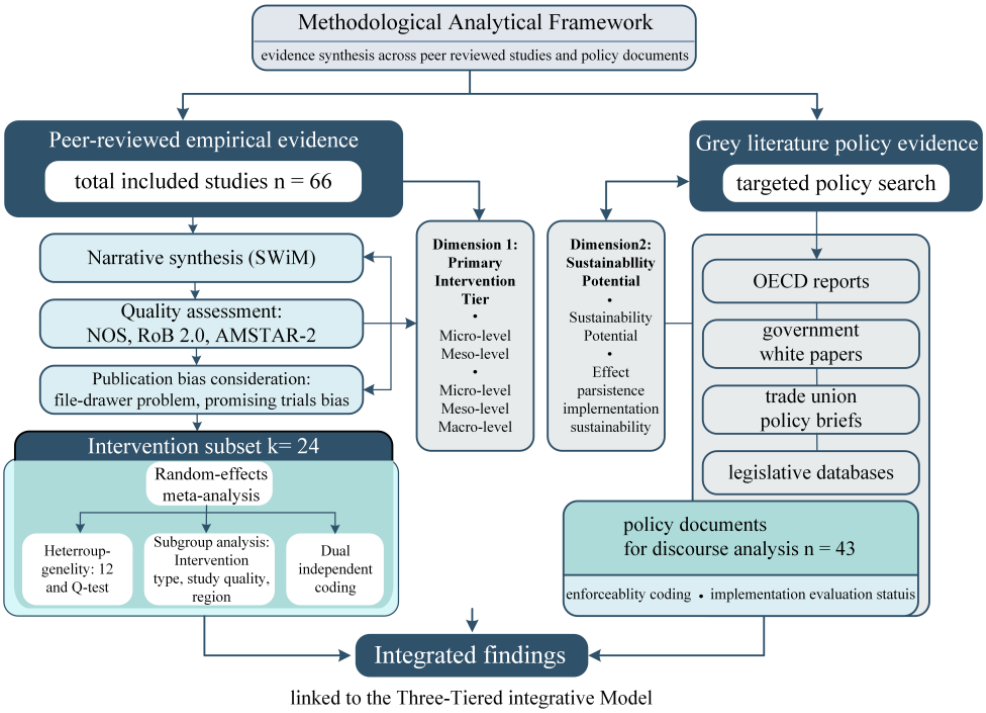

Figure 3. Analytical integration of four evidence streams.

2.4. Quality Assessment and Publication Bias

Quality was assessed using the Newcastle–Ottawa Scale (cross-sectional/longitudinal), Cochrane Risk of Bias (interventions), and AMSTAR-2 (systematic reviews); results are given in Supplementary Tables S15–S17. Quality ratings informed interpretation but did not lead to exclusion. We acknowledge positive-results bias. Rosenthal (1979) details the file drawer problem; analyses of educational RCTs identify biases in promising trials and Type M errors even in well-designed studies (Sims et al., 2023). The pooled effect size (g = 0.42) should be interpreted cautiously as a possible upper bound; the trim-and-fill-adjusted estimate (g = 0.34) is an exploratory sensitivity analysis, not a precise estimate of publication bias. Additional sensitivity analyses are given in Section 3.2 and Supplementary Table S18.

2.5. Effect-Size Synthesis and Heterogeneity

For each of the 24 intervention studies, two reviewers independently extracted bibliographic, design, sample, intervention, outcome, effect-size, and risk-of-bias variables using a pre-specified coding form; disagreements were resolved by consensus (Supplementary Tables S9 and S19). Hedges' g was computed from means and SDs, t or F statistics, or pre–post correlations; standard errors were computed using standard formulas, and 95% CIs were back-calculated as upper–lower bounds. For multiple eligible outcomes, we selected the pre-specified primary outcome or aggregated within-study effects using a random-effects model. Where means and SDs were unavailable and conversion was impossible, studies were retained in narrative synthesis but excluded from meta-analysis. Multi-arm studies were combined or the most intensive arm was selected; for clustered designs, cluster-adjusted standard errors were used where reported, otherwise a sensitivity analysis excluded non-adjusted studies. Effect directions were coded so that positive g in-

dicates an improvement in teacher mental health or well-being. Where multiple follow-ups existed, we used the longest post-intervention follow-up, except in the 2023–2026 sensitivity analysis, where the post-pandemic wave was used.

The primary model was a DerSimonian–Laird (DL) random-effects model, analyzed in R (version 4.6.1; R Core Team, 2026) using the metafor package (version 4.6-0; Viechtbauer, 2010). DL was pre-specified for comparability with prior teacher well-being syntheses (e.g., Wang et al., 2025), transparency, and protocol pre-specification. Because DL can produce anti-conservative CIs under substantial heterogeneity, we conducted pre-specified sensitivity analyses using REML, Hartung–Knapp, and Paule–Mandel estimators. Heterogeneity was substantial (I^2 = 78.6%, Q = 107.3, df = 23, $p < 0.001$); τ^2 is reported alongside I^2. Primary DL yielded τ^2 = 0.050; REML, HK, and PM yielded τ^2 = 0.060, 0.050, and 0.045, respectively. We report 95% prediction intervals (Supplementary Table S20). A pre-specified sensitivity analysis restricted to studies with data collected in 2023–2026 was conducted for prevalence estimates and the intervention meta-analysis (Section 3.1 Supplementary Table S1). Small-study effects were examined with Egger's regression and Begg's rank correlation; adjusted pooled effects used trim-and-fill and PET-PEESE. Sensitivity analyses excluded small studies (n < 50) and were restricted to studies with low risk of bias. Because these methods involve different assumptions and none of them are definitive, the results are exploratory (Supplementary Table S18).

2.6. Intervention Tier and Sustainability Coding

Two reviewers independently coded each of the 24 intervention studies on two dimensions (κ = 0.87): Dimension 1: Primary intervention tier—micro level (individual teachers' cognitive, behavioral, or emotional skills, e.g., mindfulness, CBT, and self-compassion); meso level (organizational conditions, leadership practices, and school culture, e.g., whole-school SEL and leadership coaching); and macro level (policy regimes, funding structures, and accountability frameworks). Dimension 2: Sustainability potential (exploratory), coded separately as (a) effect persistence (data-grounded dimension: present, partial, absent/unknown at $\geq$6 months) and (b) implementation sustainability (inferential dimension: low, medium, high, or based on reliance on external specialists, internal staff, or statutory policy; full anchors are provided in Supplementary Table S21). Each study was also coded for equity-relevant uptake conditions (resource requirements, high- vs. low-poverty settings, and differential subgroup effects), informing Section 3.4.

The second coding is explicitly inferential: only 7 of 24 studies reported implementation fidelity or continuation data beyond the immediate research period. Where authors commented on feasibility or adoption barriers, these notes informed coding; otherwise, coding relied on structural characteristics described in the methods. Dimension 2(b) is hypothesis-generating, not an empirical finding. The distinction between data-grounded (2a) and inferential (2b) dimensions is maintained in Table 2 and Supplementary Table S9, with the latter flagged "exploratory." A sensitivity analysis compared pooled effects for low, medium, and high sustainability (Supplementary Table S22). These analyses are exploratory and should not be interpreted as evidence that sustainability causes or moderates effects.

Table 2. Summary of intervention types, effect sizes, tier classification, and sustainability potential.

Intervention Type	Key Studies	Effect Size (g)	Primary Tier	Effect Persistence (≥6 mo)	Sustainability Potential (Exploratory)	Equity-Relevant Finding (Analytical)
CBT	Li et al. (2023); Paudel et al. (2022)	0.876	Micro	Limited (1/4)	Low	Uptake biased toward affluent settings
Mindfulness	Klingbeil and Renshaw (2018); Hidajat et al. (2023)	0.42–0.58	Micro	Partial (3/9)	Low	High dropout in low-resource schools
SEL	Oliveira et al. (2021); Cipriano et al. (2024)	0.257–0.45	Mixed	Limited	Medium	Requires admin buy-in, often absent
Positive psychology	Vo and Allen (2022)	N/A	Micro	Unknown	Low	Individualistic framing
Multi-component	Cavioni et al. (2023)	N/A (pre–post)	Meso	Unknown	Medium	No control group
Online coaching	Abdallah (2026)	N/A (*t*-test)	Micro	Unknown	Low	No workload reduction
Self-compassion	Wong et al. (2025)	N/A	Micro	Unknown	Low	May pathologize normal responses
Macro-structural reforms	None evaluated	N/A	Macro	High	High	Evidence vacuum; Nordic data suggestive

Note. Sustainability coding is exploratory and inferential. Equity findings are analytical, not normative.

2.7. Macro-Level Policy Evidence

Because macro-structural reforms are rarely evaluated via RCTs or published in peer-reviewed intervention journals, we conducted a supplementary gray literature search (January 2021–June 2026) across OECD and government reports, departmental evaluations, union briefs and collective agreements, and legislative databases (UK, US, Australia). Documents were included if they contained explicit policy provisions or implementation evaluations relevant to teacher mental health, workload, remuneration, class size, or accountability, and were available in English. Exclusions were purely advocacy statements, internal working papers, local/district documents, pre-2021 documents, and non-English documents. Two reviewers independently screened titles/abstracts and full texts; disagreements were resolved by discussion (agreement 92%; $\kappa = 0.87$), yielding 43 documents. These were analyzed narratively and coded for enforceability and independent evaluation status, not for formal risk of bias, because no validated instrument exists for this document type. They are not part of the 66-study meta-analytic corpus. Consistent with methodological pluralism (Petticrew & Roberts, 2003; Biesta, 2026), this component is treated as critical policy discourse analysis, not confirmatory causal evaluation. Full sources, search strings, and inclusion/exclusion details are provided in Supplementary Tables S8 and S10. Table 3 summarizes representative national and regional initiatives; Figure 3 illustrates the sequential workflow. The policy analysis is methodologically less rigorous than the intervention meta-analysis; policy documents were not subject to formal risk-of-bias appraisal, and the gray literature search did not follow the same PRISMA-level rigor (e.g., no dual screening with κ for all sources). This distinction is acknowledged to preempt inappropriate comparisons with the meta-analytic component.

2.8. Social Justice Analytical Framework

Social justice is treated as an analytical framework. Three commitments were specified before synthesis: (a) selection sensitivity—during screening, studies addressing equity, disparities, marginalized teacher populations, under-resourced schools, race, gender, class, or policy equity were flagged, and studies reporting subgroup or school-level socioeconomic data were prioritized where multiple eligible studies addressed the same question; (b) operationalization of disparities—equity-relevant variables were extracted where re-

ported (Supplementary Table S14); and (c) equity-weighted interpretation—pooled effects are treated as averages that may mask heterogeneity, with comparisons of high- versus low-resource settings where data permit, and assessment of whether interventions or policies target structural drivers rather than individual coping (Sections 3.4 and 4.4). This framework specifies how equity is examined, not what conclusions should be reached.

2.9. Normative Statements and Language Consistency

This review contains both empirical findings and normative arguments. To avoid conflating them, we use explicit labels throughout: [empirical finding], [theoretical interpretation], [policy implication], [normative proposal], and [speculative hypothesis]. Normative statements are identified as such and are not presented as evidence-based conclusions. Language is calibrated to the empirical weight grades in Section 4: for example, P1 (moderate) uses the "hypothesized downward pathway" and "is associated with constraints"; P3 (weak) uses the "hypothesized upward fracture" and "proposed mechanism"; P4b (speculative) uses "proposed epistemic foreclosure" and "may operate"; P2 (speculative) uses the "candidate mechanism" and "may mediate"; P5 (speculative) uses "conjecture" and "may have greater potential." Detailed language rules are included in Supplementary Table S23. Where a concept is first introduced, it is labeled with its grade; in the Results it is described as an empirical finding or theoretical interpretation, and in the Discussion it is labeled as a proposition, not a conclusion.

3. Results

Results are organized around the four evidence streams and the Three-Tier Framework (Sections 3.1–3.4); all claims follow the grading and language rules in Table 4 and Supplementary Table S11. Sixty-six studies met the inclusion criteria (Supplementary Table S6; Figure 2).

3.1. Hypothesized Downward Pathway: Macro-Level Inequities and Micro-Level Distress

Figure 4 depicts a hypothesized pathway from funding and policy decisions through organizational conditions to individual outcomes. It is an explanatory model, not a confirmed causal chain. Solid arrows denote evidence-based associations; dashed arrows denote proposed pathways; and dotted arrows denote speculative mechanisms.

Macro → meso. Evidence across 22 countries shows that national policy regimes are descriptively associated with varying prevalence rates; these patterns are hypothesis-generating only (Ma et al., 2022). Stress prevalence is 62.6% (95% Confidence Interval [CI]: 46.1–76.6) (Ma et al., 2022), with higher rates in countries with lower relative teacher salaries and higher student-to-teacher ratios (OECD, 2021a). Funding inequities—school-level free/reduced-price lunch percentage—were significantly associated with higher teacher stress and lower instructional support (Mccarthy et al., 2025; Kraft & Lyon, 2022). A 10% decrease in per-pupil expenditure predicted an approximately 7% reduction in administrative support staffing (Mccarthy et al., 2025). Pay disparities persist: Black teachers earn approximately $4400 (6%) less than White teachers, and female teachers earn about $7000 (9%) less than male teachers (Steiner et al., 2025). Nordic countries, prioritizing teacher autonomy, professional trust, and welfare support, report descriptively lower stress and attrition than the US and UK (OECD, 2021c; Sahlberg, 2021). These comparisons are heuristic only.

Meso → micro. Lack of administrative support is among the most consistent organizational predictors of burnout (Harris & Jones, 2022; Sokal et al., 2020). Increased workload, shifting responsibilities, and blurred work–home boundaries are primary meso-level stressors (Collie & Mansfield, 2022). In high-stakes testing environments, teachers

reported significantly higher moral distress—being compelled to act against professional values—than in low-stakes contexts (Sokal et al., 2020). Loh and Hu (2014) documented that examination-based accountability systems trigger emotional vulnerability by forcing teachers to prioritize metrics over meaningful pedagogy. These findings offer phenomenological plausibility, not confirmatory evidence, for moral injury (P2). They document value violation and betrayal language; they do not show mediation. No included study administered a validated moral injury instrument or used a longitudinal mediation design.

Sensitivity analysis: post-pandemic only. Using only studies with data collected in 2023–2026, prevalence estimates remained elevated: stress was 55.8% (vs. 62.6% full sample), anxiety was 32.1% (vs. 36.3%), and depression was 51.7% (vs. 59.9%). The pooled intervention effect was g = 0.45 (95% CI [0.35, 0.55]) vs. g = 0.42 for the full sample. Direction, magnitude, and subgroup pattern were materially unchanged (Supplementary Table S1).

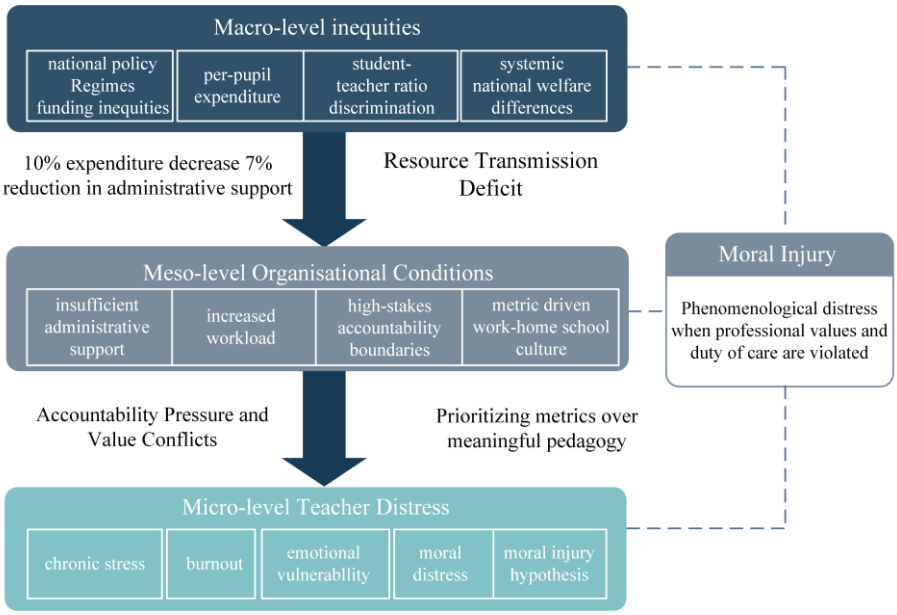

Figure 4. Proposed mechanisms of the hypothesized downward pathway.

3.2. The Great Imbalance: Why 91.7% of Interventions Target the Individual

The 2025 meta-analysis (Wang et al., 2025) found a medium positive effect (g = 0.42) on overall occupational well-being. Table 2 summarizes intervention types, effect sizes, and sustainability ratings. Sustainability coding is exploratory and inferential, based on structural characteristics described in original studies, not confirmed implementation data; only 7 of 24 studies reported implementation fidelity or continuation data. The results are hypothesis-generating and not used to estimate effectiveness. "N/A" indicates insufficient reporting for meta-analysis, not ineffectiveness. Equity-relevant findings are analytical, not normative.

Heterogeneity was substantial (I^2 = 78.6%, Q = 107.3, df = 23, $p < 0.001$). Study-level data underlying pooled estimates, as shown in Figure 5, are given in Supplementary Tables S9 and S19, making the meta-analysis fully reproducible. The primary random-effects estimate was g = 0.42 (95% CI [0.34, 0.50]). Sensitivity analyses using REML (g = 0.41, 95% CI [0.33, 0.49]), Hartung–Knapp (g = 0.41, 95% CI [0.28, 0.54]), and Paule–Mandel (g = 0.40, 95% CI [0.32, 0.48]) were comparable; the small-to-moderate positive effect remained unchanged (Supplementary Table S20).

Figure 5 presents a forest plot by intervention type. Subgroup analyses: CBT showed the largest effect (g = 0.82, 95% CI [0.61, 1.03], k = 4), but all four CBT studies were in

high-resource settings; mindfulness (k = 9) yielded g = 0.43 (95% CI [0.29, 0.57]); SEL/multi-component (k = 6) produced the smallest effect (g = 0.28, 95% CI [0.15, 0.41]). Between-group differences were significant (Q_b = 11.94, p = 0.003). Low-risk-of-bias studies (k = 9) produced g = 0.36; high risk (k = 8) produced g = 0.53 (Q_b = 4.21, p = 0.04). Egger's regression suggested possible small-study effects (t = 2.41, p = 0.025); Begg's rank correlation was p = 0.22 (τ = 0.18). The trim-and-fill-adjusted estimate was g = 0.34, suggesting possible upward bias but not a precise publication-bias estimate. Sensitivity analyses excluding small studies (N < 50) yielded g = 0.42, 95% CI [0.34, 0.50]; restricting this to low-risk-of-bias studies yielded g = 0.44, 95% CI [0.35, 0.53]; and PET-PEESE yielded g = 0.30, 95% CI [0.18, 0.42]. Direction and magnitude were comparable (Supplementary Table S18). Of the 24 intervention studies, 22 (91.7%) were micro-level, 2 (8.3%) were meso-level, and none were macro-level. This absence of published macro-level effectiveness evaluations is the central empirical manifestation of the Evidence Hierarchy Mismatch (P4). Macro-level policy initiatives exist in multiple jurisdictions (Section 3.3; Table 3) but remain underrepresented in the quantitative effectiveness literature.

Intervention Type	*k*	Pooled g [95% CI]	Weight (%)
Cognitive-Behavioral Therapy (CBT)	4	0.82 [0.61, 1.03]	18%
Mindtulness-Based Inerventions	9	0.43 [0.29, 0.57]	35%
SEL/ Multi-Component	6	0.28 [0.15,0.41]	25%
Online Coaching		0.35[0.10,0.60]	8%
Self-Compassion		0.30[0.05,0.55]	7%
Overall (random-effects model)		0.42[0.34, 0.50]	100%

0 values estimated from reported t-statistics and sample sizes

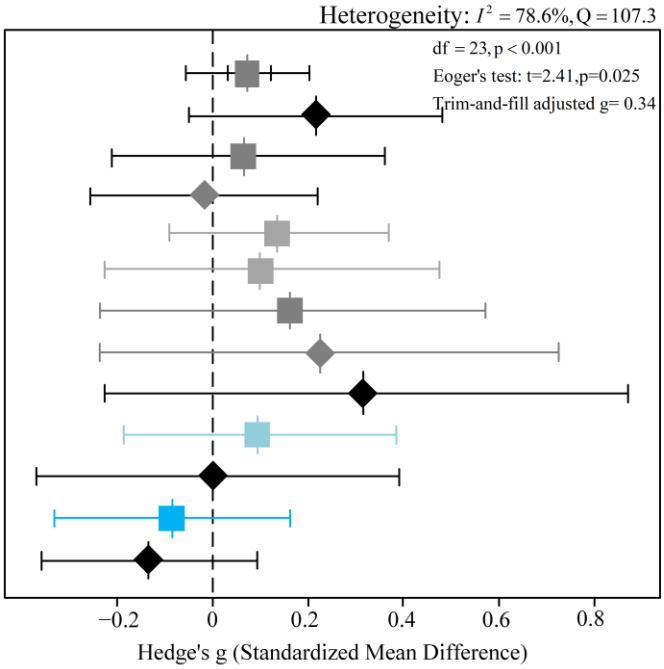

Figure 5. Forest plot of the meta-analytic effect sizes by intervention type. Note. k = 24. Data are based on Supplementary Tables S9 and S19. Sensitivity analyses are provided in Supplementary Table S20; publication-bias analyses are provided in Supplementary Table S18. Trim-and-fill estimate is exploratory.

Interpretation (theoretical). These findings are consistent with EHMT but do not establish its explanatory mechanism. The distribution documents an imbalance; it does not prove that the evidence hierarchy causes it. Alternative explanations are provided in Section 4.3.

Macro-level underrepresentation. Empirical finding: No peer-reviewed study in the identified quantitative effectiveness literature evaluated a macro-level reform. Theoretical interpretation: This is consistent with EHMT but may also reflect funding constraints, measurement difficulties, publication patterns, data access barriers, or political factors (Sections 1.2 and 4.3. The absence establishes a descriptive imbalance; it does not establish that the evidence hierarchy is the cause, nor that macro-reforms are more effective. What is known: Macro-reforms are underrepresented. What is conjectured: They may have

greater population-level potential. This remains a proposition requiring testing. Macro-reforms—salary adjustments, class-size mandates, accountability redesign—cannot be RCT-studied due to practical infeasibility (Coppe et al., 2025). Funding agencies and journals privileging RCTs may de-prioritize quasi-experimental evaluations, creating a self-reinforcing cycle: a lack of RCT evidence can lead to dismissal as "insufficiently evidenced," no funding, and rare evaluation. This is a proposed explanation, not an established causal account. Cross-national comparative evidence (Nordic vs. Anglo-American) is routinely downgraded as "correlational," confusing internal validity with policy-relevant external validity (Petticrew & Roberts, 2003). This is consistent with Nutbeam's (2004) inverse evidence law: interventions with the greatest population impact are those about which evidence is most silent.

3.3. Upward Fracture: The Broken Feedback Loop from Micro to Macro

Narrative analysis of 43 policy documents (Section 2.7) shows that while micro-level distress is pervasive—78% of teachers in one survey had thoughts of leaving (Reinke et al., 2025), and mid-career teachers are least likely to intend to stay (Longmuir, 2026)—the upward feedback loop from individual suffering to systemic reform remains fractured (P3). Table 3 summarizes international policy responses and four recurring failure mechanisms.

Table 3. Overview of international policy responses and failure mechanisms.

Country/Region	Key Policy Initiatives	Primary Tier	Enforceability	Failure Mechanism(s)	Persistent Gap
United States	RAND tracking (2021–2026)	Micro + Macro	Low	2 + 4: No binding action; no evaluation	Structural inaction despite data
United Kingdom	DfE wellbeing fund	Micro	Low	1: £10/teacher; 63% actual spend	Underfunded; policy-outcome gap
Canada	Federal pilot projects	Micro	Low	3: Workload review lost in translation	No structural follow-through
China	Health checkups; counseling	Micro	Low	3: Narrowed to individual screening	Limited evaluation
Singapore	Staff well-being committees	Meso + Micro	Medium	4: Report not public/independent	Engagement improvements only
European Union	APRES COVID; joint guidelines	Macro + Micro	Low	2: Non-binding guidelines	Variable national uptake
Nordic Countries	High autonomy; welfare; manageable workloads	Macro + Meso	High	No significant deficit	Serves as counterfactual

Note. Policy documents coded for equity targeting and enforceability; equity effects were rarely evaluated. A complete list is provided in Supplementary Table S17. Documents were classified by primary publication period (pandemic 2020–2022 vs. post-pandemic 2023–2026); analysis focuses on the post-pandemic period. The table summarizes representative initiatives, and is not exhaustive. Policy documents are interpretive sources, not evidence of peer-reviewed effectiveness.

Failure mechanism 1: Implementation deficit. The UK DfE allocated £8 million (Well-being for Education Return, 2020) and £7 million (Wellbeing for Education Recovery, 2021), primarily for training and resources rather than direct workload reduction.

Failure mechanism 2: Non-binding policy architecture. The Joint European Guidelines on Preventing Psychosocial Risks in Education (ETUCE & EFEE, 2025) and US RAND reports lack enforcement mechanisms; there is no linkage of well-being metrics to school ratings or funding. Such a "soft policy" risks re-individualizing systemic responsibility (Tsang & Wilkins, 2025).

Failure mechanism 3: Depoliticized translation. Policies are often repackaged as "teacher self-care workshops," stripping workload review and administrative support language—this is known as the "psychologized language" critique (Bekerman & Zembylas, 2025). In a Canadian case, the Educator Well-Being during Pandemic Recovery project (2023), funded with $798,500 by the Public Health Agency of Canada and implemented by

the Canadian Teachers' Federation at three pilot sites, framed its objective as "systemic" yet delivered webinars and discussion boards, re-individualizing systemic stressors.

Failure mechanism 4: Absence of evaluation. Of the eight initiatives reviewed, systematic monitoring remains limited. Singapore's MOE monitors well-being via surveys and parliamentary reporting, but no public annual implementation report was identified. Without evaluation, accountability is absent; without accountability, policy remains symbolic. Finnish provisions are embedded in legally binding collective agreements—though these specify working hours and teaching hour caps rather than class sizes, which remain a union recommendation.

The upward loop is hypothesized to be constrained by three mediators: declining union density (OECD cross-national decline from 33% in 1975 to 16% in 2018; OECD, 2019); depoliticized education discourse framing distress in individual psychological terms (Tsang & Wilkins, 2025; Wilkins et al., 2025); and media narratives attributing turnover to personal stress rather than structural conditions. This is a theoretically grounded diagnosis, not an empirically settled conclusion; relative contributions require systematic investigation. Individual distress may thus accumulate without translating into structural remedies.

3.4. Social Justice Analysis

Who is most affected from thus? Teacher distress is elevated across countries, but the burden is uneven. Teachers in high-poverty, low-resource schools report higher stress and lower administrative support (Mccarthy et al., 2025; Kraft & Lyon, 2022). Pay disparities persist by race and gender: Black teachers earned approximately $4400 (6%) less than White teachers; female teachers earn about $7000 (9%) less than male teachers (Steiner et al., 2025). These are descriptive disparities, not causal estimates.

Who benefits from these interventions? Of the 24 intervention studies, only 2 (8.3%) explicitly reported implementation in low-income or high-poverty districts. CBT studies, showing the largest pooled effect (g = 0.82), were all in high-resource settings. Mindfulness and self-compassion interventions require time, external facilitators, and release from teaching duties—resources are unevenly distributed. Voluntary micro-interventions may therefore be taken up disproportionately by teachers in better-resourced schools, potentially widening well-being gaps.

How does policy address structural inequities? The 43 policy documents included equity provisions in some jurisdictions, but enforceability varied. Nordic provisions embedded working hours and teaching-hour caps in legally binding agreements; UK and Canadian initiatives were largely non-binding and translated into individual self-care support. No policy document reported a quantitative evaluation of equity effects on teacher mental health. This is an evidence gap, not evidence of ineffectiveness.

Interpretation. Pooled effects are averages; they do not show whether interventions reduce or widen disparities. The evidence suggests micro-level interventions are least accessible to teachers facing the greatest structural burdens, while macro-level reforms with equity potential remain under-evaluated. This interpretation is analytical and provisional, not a causal conclusion.

4. Discussion

4.1. Persistence of the Crisis and the Individual-Level Paradigm

The evidence challenges any assumption that the teacher mental health crisis was transient. We treat COVID-19 as a temporal boundary condition, not a causal variable; our focus remains on structural determinants—funding inequities, accountability regimes, and evidence hierarchies—that predate and will outlast the pandemic. Teacher distress has become embedded in the profession's structure (Pressley, 2021; Collie & Mansfield,

2022; Steiner et al., 2025). Its persistence is better explained by a dominant individual-level intervention paradigm than by a lack of coping strategies.

This paradigm serves systemic responsibility diffusion (Wilkins et al., 2025): framing distress as individual failure of resilience absolves the system of responsibility for working conditions, funding inequities, and accountability pressures. Teachers' emotional experiences are socially, institutionally, and politically constructed (Karnovsky & Kelly, 2025). The "teacher wellbeing turn" (Wilkins et al., 2025) has functioned as neuropolitics, shaping teachers into "neuro-liberal workers" expected to manage systemic failures through self-regulation. Meta-analytic evidence supports this: even high-quality micro-interventions show modest, short-lived effects, dissipating within three to six months and likely inflated by publication bias. The crisis persists not because interventions are poorly designed but because the problem is misdefined.

4.2. Moral Injury as a Bridging Hypothesis

We propose moral injury as a candidate mechanism linking macro-structural constraints to micro-level distress (P2); this was advanced as testable and graded speculative. The military–education analogy and its limits are provided in Section 1.1.2 and Supplementary Note S1.

Observed. Qualitative studies document teachers using language consistent with moral injury. Sokal et al. (2020) found that 32% of Canadian teachers spontaneously invoked "betrayal" and "value violation"; Loh and Hu (2014) documented emotional vulnerability from "being forced to sacrifice pedagogical quality for metrics." These are phenomenological and do not test mediation.

Hypothesized. Resource constraints and accountability demands predict moral injury, which mediates working conditions and burnout/attrition. The proposed chain is as follows: macro-level funding cuts/accountability → meso-level resource constraints → perceived inability to provide adequate care → moral injury (guilt, shame, demoralization) → burnout/attrition. This is a theoretical model, not an empirical finding.

Not yet established. No validated moral injury instrument (e.g., MIES) has been administered to teacher samples; no included study tested moral injury as a mediator; no longitudinal three-wave mediation design exists. Validation must be teacher-specific. Moral injury is incremental to JD-R, burnout, demoralization, work stress, emotional exhaustion, and compassion fatigue (Section 1.1.1). Three validation pathways are proposed (Supplementary Table S23): (a) adapt and validate the MIES for teaching (T-MIES); (b) test discriminant and incremental validity; (c) conduct three-wave longitudinal mediation designs. Until then, moral injury serves as a theoretical device explaining why individual-level interventions may fail, reinforcing EHMT.

4.3. The Evidence Hierarchy Mismatch: Three Biases and Four Actions

Figure 6 summarizes three interlocking biases and four actions to disrupt them.

What the evidence shows [Empirical finding]: (a) 91.7% of the 24 intervention studies target the micro level; (b) 0% evaluate macro-level reforms; (c) macro-level policy initiatives exist but are substantially underrepresented in the quantitative effectiveness literature.

Theoretical interpretation: This pattern is consistent with EHMT. Three biases may sustain the imbalance:

Epistemological: The problem is the hierarchy, not RCTs. RCTs are powerful where randomization is feasible, ethical, and appropriate; the problem is that uncritical transplantation of a medical hierarchy treats RCTs as universally superior. Education deals with complex social systems where context, implementation, and value pluralism matter (Coppe

et al., 2025; Biesta, 2026). The argument presented is for methodological pluralism, not replacing RCTs.

Institutional: Funding bodies and journals often prioritize RCTs and positive results, disadvantaging quasi-experimental and policy-relevant comparative work (Parra & Edwards, 2024). D'Angelo (2025) argues that narrow definitions of "what counts" as evidence can exclude Global South perspectives. This is an institutional critique, not a claim that RCTs are politically compromised.

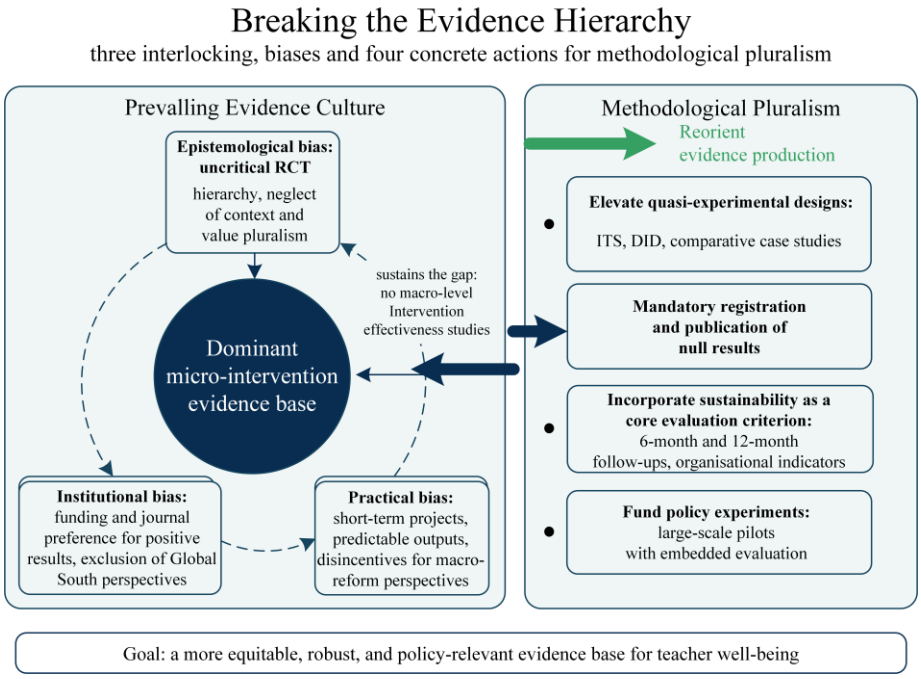

Figure 6. Breaking the evidence hierarchy.

Practical: Macro-level reforms require long-time horizons, large budgets, and cross-sectoral collaboration—they have become disincentivized in current grant-funding and promotion systems favoring short, controllable projects (Ioannidis, 2011). This affects all designs, including RCTs; it is a reason to broaden designs for policy questions.

Five competing explanations cannot be ruled out: financial limitations, measurement difficulty, publication patterns, policy data access, and political difficulties (Table S4). EHMT's distinctive prediction—that even when controlling these factors, the evidence hierarchy independently predicts lower prioritization of macro-level quasi-experimental evaluation—remains untested. EHMT is one plausible explanation among several.

Policy implications: If the imbalance is partly produced by the evidence hierarchy, policy evaluation should include quasi-experimental designs, interrupted time series, difference-in-differences, and comparative case studies. This is a policy implication derived from theoretical interpretation, not an empirical finding. It does not establish that macro-level reforms are more effective; it establishes that they are under-evaluated. Sustainability should be a core evaluation criterion.

Normative proposals: These are arguments, not empirical conclusions: (1) elevate quasi-experimental designs for macro-level policy questions (Petticrew & Roberts, 2003); (2) mandate registration and publication of null results; (3) incorporate sustainability (6- and 12-month follow-ups, organizational indicators) as a core criterion; and (4) fund large-scale policy pilots with embedded rigorous evaluation.

Methodological complementarity: RCTs suit bounded, manipulable interventions; quasi-experimental designs suit macro-level policy changes where randomization is infea-

sible, unethical, or contested. Neither is universally superior; the goal is matching designs to questions.

Known and conjectured: Known: Macro-level policy initiatives exist; they are underrepresented in the quantitative effectiveness literature; no peer-reviewed study in this review evaluated a macro-level reform using quantitative effectiveness methods. Conjectured: Macro-level reforms may have greater population-level potential, produce more sustainable effects, and address structural drivers more directly. Testing requires large-scale policy experiments with embedded quasi-experimental evaluation. Until then, superiority of macro-level interventions remains a proposition, not a conclusion.

4.4. Social Justice Implications

First, average effects can conceal inequitable effects. The pooled effect (g = 0.42) is an average across studies conducted predominantly in better-resourced settings. If effects are weaker in under-resourced schools—where baseline stress is higher and implementation resources scarcer—the average may overstate benefits for teachers who need them most. This is a hypothesis, not a confirmed finding.

Second, intervention design and delivery have equity consequences. Interventions requiring external facilitators, release time, or recurring budgets are less sustainable and accessible in high-poverty schools. Sustainability coding (Table 2) is therefore an equity issue. Statutory-policy interventions (staffing ratios and workload caps) may have greater equity potential because they do not depend on individual teacher uptake, but this remains untested.

Third, policy analysis must ask who is targeted and who is reached. Universal policies may be regressive if uptake is higher in well-resourced settings; targeted policies may reduce gaps but risk stigmatization. The reviewed policy documents rarely reported equity effects. Future evaluation should report differential effects by school poverty, race/ethnicity, gender, and career stage.

The unequal distribution of occupational stress reflects broader educational inequality. Persistent disparities by race and gender—with Black teachers more likely to intend to leave and female teachers reporting higher stress levels and burnout (Steiner et al., 2025)—demand structural policy attention. Over-reliance on micro-level interventions exacerbates inequities: teachers in under-resourced schools have less time, fewer substitutes, and higher baseline stress, making voluntary mindfulness programs difficult to access. A socially just approach requires the following: (a) targeting resources to the most affected schools and populations; (b) addressing structural determinants including workload, pay, status, and autonomy; and (c) centering intervention design and policy advocacy around teachers of color, women, and those in low-wealth schools.

Cross-national comparative evidence. The Nordic–Anglo-American comparison is descriptive and hypothesis-generating (Sahlberg, 2021; OECD, 2021b). It is confounded by historical trajectories, cultural values, unionization histories, and welfare regimes. No causal claims are derived. We use this contrast to illustrate plausible macro-level policy variation and to motivate future within-country quasi-experimental studies.

4.5. Empirical Weight of the Theoretical Propositions

This section integrates the four evidence streams. Table 4 grades each proposition; Supplementary Table S13 maps each stream to its integration point.

Table 4. Empirical weight of the core theoretical propositions (P1–P5).

Proposition	Summary of Claim	Evidence Currently Available	Empirical Weight	Key Sources/Identified Gap
P1	Macro-level funding inequities predict meso-level resource constraints, which predict micro-level distress (hypothesized downward pathway)	Cross-national comparative studies; multi-level cross-sectional studies; funding–staffing regressions. Individual links supported; full mediation chain not tested longitudinally or causally.	Moderate (downgraded from strong)	OECD (2021a); Mccarthy et al. (2025); Kraft and Lyon (2022); no longitudinal or quasi-experimental test of full macro → meso → micro mediation chain identified
P2	Resource constraints and accountability pressures predict moral injury, which mediates working conditions and burnout/attrition	Observed: qualitative reports of betrayal/value violation (Sokal et al., 2020; Loh & Hu, 2014). Hypothesis: mediation untested; no validated scale-based studies; no longitudinal mediation design.	Speculative	Sokal et al. (2020); Loh and Hu (2014); no T-MIES or longitudinal mediation study to date
P3	Upward feedback loop (micro → macro) moderated by union density, media framing, and political receptivity	Cross-national institutional comparisons; policy discourse analyses	Weak (directional)	Sahlberg (2021); Tsang and Wilkins (2025); no direct moderator tests available
P4a	The published quantitative effectiveness literature is skewed toward micro-level interventions and away from macro-level reforms (descriptive)	Tier distribution: 91.7% micro and 0% macro	Moderate (descriptive)	Table 2, this review; Nutbeam (2004)
P4b	This skew is produced primarily by the evidence hierarchy (epistemological, institutional, and practical biases) (explanatory)	No direct test; competing explanations not ruled out	Speculative (explanatory)	Coppe et al. (2025); Parra and Edwards (2024); Petticrew and Roberts (2003); competing explanations in Sections 1.2 and 4.3.
P5	Macro-level structural reforms may produce greater population-level reductions in teacher distress than micro-level interventions	No quantitative effectiveness study identified; derived from theoretical reasoning and policy analysis	Speculative	No peer-reviewed macro-level effectiveness evaluation identified; requires policy experiments with embedded quasi-experimental evaluation

Note. Grading reflects evidence support, not normative importance; social justice informs interpretation without altering grades. P4a is descriptive; P4b is explanatory; P2 distinguishes observation from hypothesized mediation. Strong = multiple consistent quantitative studies or meta-analyses; moderate = consistent links but limited causal or longitudinal testing (P1, since the full macro → meso → micro chain is untested); weak = sparse or indirect evidence; speculative = untested. No proposition establishes a causal chain.

P1 has the strongest empirical anchoring but is graded moderate: individual links are supported, but the full mediation chain is untested. P2 remains at the hypothesis stage. P3 rests on institutional inference. P4a is supported by the tier distribution; P4b remains speculative. P5 is graded speculative: the review establishes underrepresentation, not superiority.

Theoretical contributions. Three advances emerge. First, the Three-Tier Integrative Framework specifies a proposed vertical pathway (macro → meso → micro) as well as conditions for upward feedback (micro → macro), extending JD-R with vertical stratification and social justice criteria. Second, EHMT specifies institutional mechanisms—such as epistemological transplantation, funding incentives, and practical infeasibility—producing evidence-base distortion, and introduces "proposed epistemic foreclosure." Third, the moral injury hypothesis provides a candidate psychological mechanism bridging macro-structural and micro-experiential levels. Collectively, these contributions shift the theoretical center of gravity from individual resilience to structural explanation, from intervention efficacy to evidence architecture, and from descriptive prevalence to explanatory mechanisms.

4.6. Limitations

Search and scope. The search was limited to English-language publications, and the post-pandemic focus may under-capture long-term effects. Pre-2021 publications were not systematically captured. The inclusion of diverse designs limits quantitative conclusions.

Design coverage. Restricting meta-analysis to intervention studies with synthesizable effect sizes (k = 24) excluded qualitative and mixed-methods studies that could illuminate organizational and systemic change. Future reviews should synthesize this complementary evidence.

Policy analysis asymmetry. The supplementary gray literature search did not follow PRISMA-level rigor (no dual screening with κ and no formal quality appraisal). Policy findings are interpretive, not evidentiary, and should not be treated as equivalent to meta-analytic findings.

Moral injury hypothesis. No validated instrument has been administered to teacher samples, and no longitudinal mediation study exists. Teacher-specific validation is required. Moral injury is complementary and speculative, and not a replacement for established constructs.

Publication bias. The trim-and-fill estimate (g = 0.34) is exploratory, and not precise. Publication bias cannot be eliminated, and the pooled effect should be interpreted with caution.

Social justice analysis. The equity-weighted interpretation is limited by available equity data and is partly inferential. Future reviews require equity-relevant reporting.

5. Conclusions

This review documents persistent, elevated teacher mental health difficulties in the post-pandemic era, extending from individual suffering to educational quality and equity. Interventions remain overwhelmingly micro-level, while macro-structural reforms are substantially underrepresented in the identified quantitative effectiveness literature—not because they are absent in practice, but because the dominant evidence hierarchy privileges designs often infeasible for system-level policy evaluation. Cross-national policy comparisons are heuristic illustrations, not causal evidence.

The core challenge is a paradigm lag: the field prioritizes individual-level methodological conventions for structurally produced problems. As long as RCTs are treated as the only rigorous design, macro-level reforms may remain under-evidenced and under-funded. We do not conclude that macro-level reforms are more effective; their comparative effectiveness remains an open empirical question. The proposition that they may have greater population-level potential is a conjecture requiring testing. The argument is not against RCTs but for matching designs to questions and for methodological pluralism.

Normatively, we call for a paradigm shift from individual remediation to systemic reform, and we propose methodological pluralism, mandatory null-results publication, and sustainability as a core evaluation criterion. These are arguments about what should be done, not empirical conclusions. The Three-Tier Framework and Evidence Hierarchy Mismatch Theory provide theoretical scaffolding; moral injury remains a candidate bridging mechanism awaiting empirical validation.

Education sciences is well positioned to lead this shift by welcoming diverse methodologies, prioritizing structural analyses, and reframing teacher well-being from an individual luxury to a systemic necessity.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/educsci16101642/s1. Table S1: Post-Pandemic (2023–2026) Sensitivity Analyses; Table S2: Component–Method–Evidence Standard–Interpretive Status Mapping; Table S3: Moral injury in relation to established teacher well-being constructs; Table S4: Summary of Five Competing Explanations; Table S5: Complete Search Strategies for All Databases; Table S6: Characteristics of Included Studies (k = 66); Table S7: PRISMA 2020 Checklist; Table S8: Excluded Studies with Reasons (Representative List); Table S9: Complete List of 24 Intervention Effectiveness Studies Included in the Meta-Analysis (with Tier Classification and Sustainability Coding); Table S10: Data Extraction Form Template; Table S11: Equity-relevant variables extracted from included studies; Table S12: Language Rules and Wording Calibration; Table S13: Integration Matrix of Four Evidence Streams; Table S14: Equity-Relevant Variables Extraction Table; Table S15: Quality Assessment Ratings for Prevalence Studies (Newcastle-Ottawa Scale) Table; Table S16: Quality Assessment Ratings for Intervention Studies (Cochrane RoB 2.0); Table S17: Quality Assessment Ratings for Policy Reports (AMSTAR-2); Table S18: Publication-Bias and Additional Sensitivity Analyses; Table S19: Effect-Size Extraction and Coding for Intervention Effectiveness Studies (k = 24); Table S20: 95% Prediction Intervals and Heterogeneity Sensitivity Analyses; Table S21: Sustainability Coding Anchors; Table S22: Sensitivity Analysis of Pooled Effects by Sustainability Level; Table S23: Validation Pathways for the Teacher Moral Injury Events Scale (T-MIES). Supplementary Note S1: The Military–Education Analogy for Moral Injury: Scope, Limitations, and Risks of Conceptual Overreach.

Author Contributions: Conceptualization, Z.W. and B.L.; methodology, Z.W. and B.L.; validation, Z.W. and B.L.; formal analysis, Z.W.; investigation, Z.W.; resources, Z.W. and B.L.; data curation, Z.W. and B.L.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W. and B.L.; visualization, Z.W. and B.L.; supervision, Z.W. and B.L.; project administration, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding; the APC was funded by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data analyzed in this study—the full coding matrix of the peer reviewed and practitioner/policy corpora—are contained within the article. The primary sources analyzed are publicly available and listed in the reference list. No new primary data were created.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

COVID-19	Coronavirus Disease 2019
EHMT	Evidence Hierarchy Mismatch Theory
JD-R	Job Demands-Resources Model
P1–P5	Propositions 1 to 5
RQ1–RQ3	Research Questions 1 to 3
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomized Controlled Trial
CBT	Cognitive-Behavioral Therapy
SEL	Social and Emotional Learning
SWiM	Synthesis Without Meta-analysis
AMSTAR-2	A Measurement Tool to Assess Systematic Reviews 2
NOS	Newcastle–Ottawa Scale
RoB	Risk of Bias
CI	Confidence Interval

df	Degrees of Freedom
OECD	Organisation for Economic Co-operation and Development
ETUCE	European Trade Union Committee for Education
EFEE	European Federation of Education Employers
DfE	Department for Education (UK)
ED	Department of Education (US)
DET	Department of Education and Training (Australia)
MOE	Ministry of Education (Singapore)
NEA	National Education Association (US)
MIES	Moral Injury Events Scale
T-MIES	Teacher Moral Injury Events Scale

References

Abdallah, A. K. (2026). Effects of online content-focused coaching on reducing the risk of professional burnout among elementary school teachers: The example of educational leadership in the United Arab Emirates. *The Journal of General Psychology*, *153*(3), 342–365. [CrossRef] [PubMed]

Acuña, F. (2024). Governing teachers' subjectivity in neoliberal times: The fabrication of the bonsai teacher. *Journal of Education Policy*, *39*(2), 171–190. [CrossRef]

Aziku, M., & Zhang, B. (2024). Systematic review of teacher well-being research during the COVID-19 pandemic. *Frontiers in Psychology*, *15*, 1427979. [CrossRef] [PubMed]

Bakker, A. B., & Demerouti, E. (2017). Job demands–resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, *22*(3), 273–285. [CrossRef] [PubMed]

Bekerman, Z., & Zembylas, M. (2025). Reflecting on metaphors and the possibilities of 'language change' in teaching and teacher education. *European Journal of Teacher Education*, *48*(4), 874–890. [CrossRef]

Berger, E., Reupert, A., Campbell, T. C. H., Morris, Z., Hammer, M., Diamond, Z., Hine, R., Patrick, P., & Fathers, C. (2022). A systematic review of evidence-based wellbeing initiatives for schoolteachers and early childhood educators. *Educational Psychology Review*, *34*(4), 2919–2969. [CrossRef]

Biesta, G. (2026). Has research on "what works" in education been looking at the wrong interventions? Revisiting "why 'what works' won't work". *Educational Theory*, *76*(1), 16–24. [CrossRef]

Cavioni, V., Grazzani, I., Ornaghi, V., Agliati, A., Gandellini, S., Cefai, C., Camilleri, L., Bartolo, P., Vorkapić, S. T., Golob, L., Poulou, M., Martinsone, B., Supe, I., Simoes, C., Lebre, P., Colomeischi, A., Rusu, P., Acostoaie, L., Vintur, T., & Conte, E. (2023). A multi-component curriculum to promote teachers' mental health: Findings from the PROMEHS program. *International Journal of Emotional Education*, *15*(1), 34–52. [CrossRef]

Cipriano, C., Ha, C., Wood, M., Sehgal, K., Ahmad, E., & McCarthy, M. F. (2024). A systematic review and meta-analysis of the effects of universal school-based SEL programs in the United States: Considerations for marginalized students. *Social and Emotional Learning: Research, Practice, and Policy*, *3*, 100029. [CrossRef]

Collie, R. J., & Mansfield, C. F. (2022). Teacher and school stress profiles: A multilevel examination and associations with work-related outcomes. *Teaching and Teacher Education*, *116*, 103759. [CrossRef]

Coppe, T., Hoekstra, R., Helgetun, J. B., Schwarzbach, N. R., & van der Velden, M. (2025). Evidence-based research in education: The questionable epistemic status of RCTs. *Critical Studies in Education*, *67*, 525–541. [CrossRef]

D'Angelo, S. (2025). *Widening the evidence base in education: The urgent need to expand 'what counts' to deliver better policy making*. NORRAG Blog. Available online: https://www.norrag.org/widening-the-evidence-base-in-education-the-urgent-need-to-expand-what-counts-to-deliver-better-policy-making (accessed on 17 July 2026).

ETUCE & EFEE. (2025). *Joint European guidelines on preventing psychosocial risks in education*. Available online: https://etuce.org/en/item/4462:joint-european-guidelines-on-preventing-psychosocial-risks-in-education-2025 (accessed on 18 July 2026).

Harris, A., & Jones, M. (2022). Leading during a pandemic—What the evidence tells us. *School Leadership & Management*, *42*(2), 105–109. [CrossRef]

Hidajat, T. J., Edwards, E. J., Wood, R., & Campbell, M. (2023). Mindfulness-based interventions for stress and burnout in teachers: A systematic review. *Teaching and Teacher Education*, *134*, 104303. [CrossRef]

Ioannidis, J. P. A. (2011). Fund people not projects. *Nature*, *477*(7366), 529–531. [CrossRef] [PubMed]

Karnovsky, S., & Kelly, N. (Eds.). (2025). *Teachers' emotional experiences: Towards a new emotional discourse* (Palgrave critical perspectives on schooling, teachers and teaching). Palgrave Macmillan. [CrossRef]

Klingbeil, D. A., & Renshaw, T. L. (2018). Mindfulness-based interventions for teachers: A meta-analysis of the emerging evidence base. *School Psychology Quarterly*, *33*(4), 501–511. [CrossRef] [PubMed]

Kraft, M. A., & Lyon, M. A. (2022). *The rise and fall of the teaching profession: Prestige, interest, preparation, and satisfaction over the last half century. EdWorkingPaper No. 22-679*. Annenberg Institute for School Reform at Brown University. Available online: https://www.proquest.com/reports/rise-fall-teaching-profession-prestige-interest/docview/2860865266/se-2 (accessed on 15 July 2026).

Li, J., Xue, E., & He, Y. (2023). Investigating the effect of cognitive–behavioral, mindful-based, emotional-based intervention and professional training on teachers' job burnout: A meta-analysis. *Behavioral Sciences*, *13*(10), 803. [CrossRef] [PubMed]

Loh, J., & Hu, G. (2014). Subdued by the system: Neoliberalism and the beginning teacher. *Teaching and Teacher Education*, *41*, 13–21. [CrossRef]

Longmuir, F. (2026). Burnt-out and demoralised: Mid-career attrition from the teaching workforce. *Teachers and Teaching*, 1–17. [CrossRef]

Ma, K., Liang, L., Chutiyami, M., Nicoll, S., Khaerudin, T., & Ha, X. V. (2022). COVID-19 pandemic-related anxiety, stress, and depression among teachers: A systematic review and meta-analysis. *Work*, *73*(1), 3–27. [CrossRef] [PubMed]

Mccarthy, C. J., Lambert, R. G., Yijie, T., & Syed, R. M. (2025). Teacher appraisals of occupational wellbeing before and after COVID-19. *Journal of Education, Society & Multiculturalism*, *6*(2), 58–77. [CrossRef]

Nutbeam, D. (2004). Getting evidence into policy and practice to address health inequalities. *Health Promotion International*, *19*(2), 137. [CrossRef] [PubMed]

Oliveira, S., Roberto, M. S., Veiga-Simão, A. M., & Marques-Pinto, A. (2021). A meta-analysis of the impact of social and emotional learning interventions on teachers' burnout symptoms. *Educational Psychology Review*, *33*(4), 1779–1808. [CrossRef]

Organisation for Economic Co-Operation and Development [OECD]. (2019). *Negotiating our way up: Collective bargaining in a changing world of work*. OECD Publishing. [CrossRef]

Organisation for Economic Co-Operation and Development [OECD]. (2021a). *Education at a glance 2021: OECD indicators*. OECD Publishing. [CrossRef]

Organisation for Economic Co-Operation and Development [OECD]. (2021b). *Teachers getting the best out of their students: From primary to upper secondary education*. TALIS, OECD Publishing. [CrossRef]

Organisation for Economic Co-Operation and Development [OECD]. (2021c, December 7). *Building teachers' well-being from primary to upper secondary education* (Teaching in Focus, No. 42). OECD Publishing. [CrossRef]

Ozamiz-Etxebarria, N., Mondragon, N. I., Bueno-Notivol, J., Pérez-Moreno, M., & Santabárbara, J. (2021). Prevalence of anxiety, depression, and stress among teachers during the COVID-19 pandemic: A rapid systematic review with meta-analysis. *Brain Sciences*, *11*(9), 1172. [CrossRef] [PubMed]

Parra, J. D., & Edwards, D. B. (2024). Challenging the gold standard consensus: Randomised Controlled Trials (RCTs) and their pitfalls in evidence-based education. *Critical Studies in Education*, *65*(5), 513–530. [CrossRef]

Paudel, N. R., Adhikari, B. A., Prakash, K. C., Kyrönlahti, S., Nygård, C. H., & Neupane, S. (2022). Effectiveness of interventions on the stress management of schoolteachers: A systematic review and meta-analysis. *Occupational and Environmental Medicine*, *79*(7), 477–485. [CrossRef] [PubMed]

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiology and Community Health*, *57*(7), 527–529. [CrossRef] [PubMed]

Pressley, T. (2021). Factors contributing to teacher burnout during COVID-19. *Educational Researcher*, *50*(5), 325–327. [CrossRef]

R Core Team. (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available online: https://www.R-project.org (accessed on 18 July 2026).

Reinke, W. M., Herman, K. C., Stormont, M., & Ghasemi, F. (2025). Teacher stress, coping, burnout, and plans to leave the field: A post-pandemic survey. *School Mental Health*, *17*(1), 32–44. [CrossRef]

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. [CrossRef]

Sahlberg, P. (2021). *Finnish lessons 3.0: What can the world learn from educational change in Finland?* (3rd ed.). Teachers College Press.

Silva, D. F. O., Cobucci, R. N., Lima, S. C. V. C., & de Andrade, F. B. (2021). Prevalence of anxiety, depression, and stress among teachers during the COVID-19 pandemic: A PRISMA-compliant systematic review. *Medicine*, *100*(44), e27684. [CrossRef] [PubMed]

Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2023). Quantifying "promising trials bias" in randomized controlled trials in education. *Journal of Research on Educational Effectiveness*, *16*(4), 663–680. [CrossRef]

Sokal, L. J., Eblie Trudel, L. G., & Babb, J. C. (2020). Supporting teachers in times of change: The job demands-resources model and teacher burnout during the COVID-19 pandemic. *International Journal of Contemporary Education*, *3*(2), 67–74. [CrossRef]

Steiner, E. D., Phoebe, R. L., Doan, S., & Woo, A. (2025). *Teacher well-being, pay, and intentions to leave in 2025: Findings from the state of the American teacher survey* (American Educator Panels. Data Note. RR-A1108-16). RAND Corporation. [CrossRef]

Tsang, K. K., & Wilkins, A. W. (2025). The politics of teacher wellbeing: 'Sung baang', neoliberalism and power struggles in Hong Kong. *Globalisation, Societies and Education*, 1–15. [CrossRef]

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. [CrossRef]

Vo, D. T., & Allen, K.-A. (2022). A systematic review of school-based positive psychology interventions to foster teacher wellbeing. *Teachers and Teaching: Theory and Practice*, *28*(8), 964–999. [CrossRef]

Wang, H., Cha, Y., Wen, Y., & Lim, C. (2025). Interventions that promote teachers' occupational well-being: A meta-analytic investigation of their effectiveness. *Review of Educational Research*, *96*(5), 1597–1639. [CrossRef]

Wilkins, A. W., Olmedo, A., Gobby, B., & Karnovsky, S. (2025). The teacher wellbeing turn: Neuropolitics, education and the psy-complex. *Critical Studies in Education*, 1–21. [CrossRef]

Wong, M. Y. C., Fung, H. W., Yuan, G. F., & Tao, S. Y. (2025). The role of self-compassion in teachers' well-being: A systematic review and meta-analysis. *International Journal of Psychology*, *60*(5), e70098. [CrossRef] [PubMed]

Zembylas, M. (2021). Against the psychologization of "resilience": Towards an onto-political theorization of the concept and its implications for higher education. *Studies in Higher Education*, *46*(9), 1966–1977. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.