Measuring moral distress and moral injury

DOI:
10.1016/j.cpr.2023.102377

Publication date:
2024

Clinical Psychology Review 108 (2024) 102377

Contents lists available at ScienceDirect

Clinical Psychology Review

journal homepage: www.elsevier.com/locate/clinpsychrev

Review

Measuring moral distress and moral injury: A systematic review and content analysis of existing scales

Stephanie A. Houle [a,b], Natalie Ein [a,c], Julia Gervasio [a], Rachel A. Plouffe [a,d], Brett T. Litz [e,f,g], R. Nicholas Carleton [h], Kevin T. Hansen [a], Jenny J.W. Liu [a,c], Andrea R. Ashbaugh [i], Walter Callaghan [j], Megan M. Thompson [k], Bethany Easterbrook [a,l], Lorraine Smith-MacDonald [m], Sara Rodrigues [n], Stéphanie A.H. Bélanger [o], Katherine Bright [p], Ruth A. Lanius [c], Clara Baker [a], William Younger [a], Suzette Bremault-Phillips [q], Fardous Hosseiny [n], J. Don Richardson [a,c,r,s], Anthony Nazarov [a,c,s,*], the Atlas Institute Moral Injury Research Community of Practice

[a] MacDonald Franklin OSI Research Centre, Lawson Health Research Institute, London, Canada
[b] Research Directorate, Veterans Affairs Canada, Charlottetown, Canada
[c] Department of Psychiatry, Western University, London, Canada
[d] Department of Psychology, University of Dundee, Dundee, UK
[e] Department of Psychiatry, Boston University, Boston, USA
[f] Massachusetts Veterans Epidemiology Research and Information Center, VA Boston Healthcare System, Boston, USA
[g] Department of Psychological and Brain Sciences, Boston University, Boston, USA
[h] Department of Psychology, University of Regina, Regina, Canada
[i] School of Psychology, University of Ottawa, Ottawa, Canada
[j] Department of Anthropology, University of Toronto, Toronto, Canada
[k] Defence Research and Development Canada, Toronto, Canada
[l] Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, Canada
[m] St. Stephens College, University of Alberta, Edmonton, Canada
[n] The Atlas Institute for Veterans and Families, Ottawa, Canada
[o] Royal Military College of Canada, Kingston, Canada
[p] Nursing and Midwifery, Mount Royal University, Calgary, Canada
[q] Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, Canada
[r] St. Joseph's Operational Stress Injury Clinic, St. Joseph's Health Care London, London, Canada
[s] Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada

ARTICLE INFO

Keywords:
Moral distress
Moral injury
Psychometrics
Scale development
Measurement

ABSTRACT

Background: Moral distress (MD) and moral injury (MI) are related constructs describing the negative consequences of morally challenging stressors. Despite growing support for the clinical relevance of these constructs, ongoing challenges regarding measurement quality risk limiting research and clinical advances. This study summarizes the nature, quality, and utility of existing MD and MI scales, and provides recommendations for future use.

Method: We identified psychometric studies describing the development or validation of MD or MI scales and extracted information on methodological and psychometric qualities. Content analyses identified specific outcomes measured by each scale.

Results: We reviewed 77 studies representing 42 unique scales. The quality of psychometric approaches varied greatly across studies, and most failed to examine convergent and divergent validity. Content analyses indicated most scales measure exposures to potential moral stressors and outcomes together, with relatively few measuring only exposures ($n = 3$) or outcomes ($n = 7$). Scales using the term MD typically assess general distress. Scales using the term MI typically assess several specific outcomes.

* Corresponding author at: MacDonald Franklin OSI Research Centre, Lawson Health Research Institute, London, Canada.
E-mail address: anthony.nazarov@sjhc.london.on.ca (A. Nazarov).

https://doi.org/10.1016/j.cpr.2023.102377
Received 31 August 2023; Received in revised form 3 November 2023; Accepted 21 December 2023
Available online 27 December 2023

Conclusions: Results show how the terms MD and MI are applied in research. Several scales were identified as appropriate for research and clinical use. Recommendations for the application, development, and validation of MD and MI scales are provided.

1. Introduction

Research and clinical interest in the emotional, psychosocial, and health impacts of moral stressors is growing substantially, prompting a need for clarity in the operationalization and measurement of moral stressors and their outcomes. Specifically, the intended meaning of the terms moral distress (MD) and moral injury (MI) are becoming increasingly blurred, and improper measurement of the nature and impact of moral stressors risks affecting the validity of inferences that can be drawn from research (Litz & Kerig, 2019; Plouffe et al., 2021; Plouffe et al., 2021). As such, the current review focuses on describing the nature, quality, and utility of existing measures of MD and MI, and to provide guidance relative to the application of these constructs and related instruments in research and clinical practice.

MD and MI are related constructs commonly used in occupational health contexts to describe the effects of experiences that violate one's moral values and beliefs (Hall, Everson, Billingsley, & Miller, 2021; Lamiani, Borghi, & Argentero, 2017). MD is derived primarily from the nursing and healthcare field and is commonly defined as arising "when one knows the right thing to do, but institutional constraints make it nearly impossible to pursue the right course of action" (Jameton, 1984, p. 6). MI, a term that chiefly originated in military behavioural health contexts, has been defined as "the lasting psychological, biological, spiritual, behavioral, and social impact of perpetrating, failing to prevent, or bearing witness to acts that transgress deeply held moral beliefs and expectations" (Litz et al., 2009, p. 700). The terms "distress" and "injury" may also be understood to represent different degrees of harm experienced across a continuum of moral stressors and outcomes (Litz & Kerig, 2019), with "distress" describing moderate impacts from relatively common moral stressors (e.g., being lied to) and "injury" describing a more severe and functionally impairing outcome in response to high stakes and higher magnitude events (e.g., bearing witness to human cruelty). While the reality that institutional and situational constraints can present moral challenges has long been acknowledged in both healthcare and the military, MD as operationalized in healthcare has a more substantial research history, with most evidence investigating associations between exposure to moral stressors and organizational factors (e.g., negative ethical climate) and occupational functioning (e.g., burnout, job satisfaction; Lamiani et al., 2017; Oh & Gastmans, 2015). Still, conceptual fuzziness of MD has long been noted (Dean, Talbot, & Caplan, 2020; Deschenes, Gagnon, Park, & Kunyk, 2020; Epstein & Hamric, 2009; McCarthy & Deady, 2008; Morley, Ives, Bradbury-Jones, & Irvine, 2019; Ritchie, 2019), perhaps explaining why, as military MI research began to describe different potentially morally injurious events (PMIEs) and their consequences, and as the COVID-19 pandemic highlighted various moral stressors among front-line workers (Plouffe, Easterbrook, et al., 2021, Plouffe, Nazarov, et al., 2021), convergence between MD and MI research accelerated. Measures of MI developed for the military context were quickly adapted to health care (Mantri, Lawson, Wang, & Koenig, 2020; Morris, Webb, Trundle, & Caetano, 2022), and associations among MI and mental health in this context began to show associations with psychiatric outcomes in addition to occupational functioning (Mantri, Lawson, Wang, & Koenig, 2021; Plouffe, Easterbrook, et al., 2021; Plouffe, Nazarov, et al., 2021). Unfortunately, convergence between MD and MI remains to be properly elaborated, with ongoing measurement issues posing further risks to conceptual clarity.

Much like exposure to potentially traumatic events, exposure to moral stressors, regardless of intensity, does not invariably lead to problematic outcomes; however, many MD and MI scales fail to separate the assessment of exposure to moral stressors from potentially harmful sequelae. The conflation of exposure and outcome has unfortunately impeded our understanding of the risks associated with exposure and hindered the identification of effective mitigation and intervention strategies (Kolbe & de Melo-Martin, 2022; Morley et al., 2019; Plouffe, Easterbrook, et al., 2021; Plouffe, Nazarov, et al., 2021). For example, the Moral Injury Events Scale (MIES; Nash et al., 2013), the first psychometric scale designed to purportedly measure MI in a military population, has been used extensively in research and clinical settings. Certain MIES items, however, assess event exposure (e.g., "I acted in ways that violated my own moral code or values") while others imply an outcome without clearly specifying its nature (e.g., "I feel betrayed by leaders I once trusted"). Some have therefore treated the MIES as a measure PMIE exposure (e.g., Zerach, Ben-Yehuda, & Levi-Belz, 2023), while others have conceptualized the scale as reflective of an outcome (e.g., Nillni et al., 2020). Similarly, the widely used Moral Distress Scale-Revised (MDS-R; Hamric, Borchers, & Epstein, 2012) and Measure of Moral Distress for Health Professionals (MMD-HP; Epstein, Whitehead, Prompahakul, Thacker, & Hamric, 2019) both combine the evaluation of exposure frequencies and associated distress levels for highly specific healthcare scenarios (e.g., committing a medication error). This dual assessment approach obscures our ability to determine which moral stressor(s) putatively lead to outcomes, which has impeded targeted interventions to mitigate MD in the workplace (Kolbe & de Melo-Martin, 2022).

Failing to disaggregate moral stressor exposures from outcomes is representative of a broader psychometric problem regarding the proper application of measurement models in scale design and validation (Coltman, Devinney, Midgley, & Venaik, 2008). There are two types of measurement models typically used in developing psychometric scales – formative and reflective. Formative models contain items expected to compose a construct without necessarily being inter-correlated (e.g., an assessment for trauma exposure that includes items about natural disasters and sexual assault). Reflective models include items expected to each describe a particular latent construct (e.g., items assessing self-critical evaluations and excessively high standards for one's behaviour as reflective of perfectionism). Unfortunately, recent evaluations of psychometric scales of MD have not accounted for this important distinction, further perpetuating measurement issues in this area (Giannetta et al., 2020). Applying appropriate measurement models and tools is essential for construct validation and clinical utility when assessing stressor-related problems (Karstoft & Armour, 2023; Kolbe & de Melo-Martin, 2022; Litz & Kerig, 2019). For example, per leading taxonomies' caseness rules, assessing formative (i.e., exposure to severe high magnitude life stressors) and reflective (i.e., symptom presentation) components is required to inform diagnosis and treatment of posttraumatic stress disorder (PTSD; American Psychiatric Association, 2022; World Health Organization, 2019–2021). Leading conceptualizations of MD and MI similarly maintain that there can be no outcome independent of a moral stressor (Corley, 2002; Farnsworth, Drescher, Evans, & Walser, 2017; Litz & Kerig, 2019). Research results to date support this conceptualization, demonstrating associations between moral stressor exposure and symptoms of anxiety-, depressive-, and trauma-related disorders, as well as suicidality (Easterbrook et al., 2023; Griffin et al., 2019; Hall et al., 2021; Nazarov, Fikretoglu, Liu, Thompson, & Zamorski, 2018; Riedel, Kreh, Kulcar, Lieber, & Juen, 2022).

Still, a paradigmatic approach to the measurement of MD and MI is both lacking and needed. Despite growing support for the clinical relevance of these constructs, the proliferation of measures in this area coupled with ongoing challenges with measurement quality risk limiting

potential research and clinical advances. We addressed this issue by conducting a comprehensive review of the nature, quality, and utility of extant MD and MI measures, and provide recommendations for assessing MD and MI in research, occupational, and clinical practice.

2. Methods

The current review was conducted according to Cochrane's guidelines (Higgins et al., 2022) and adapted criteria from the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN; Mokkink et al., 2018). The Cochrane guidelines include a search strategy across multiple databases for published and unpublished studies, two levels of screening (title and abstract and full text) against inclusion and exclusion criteria, resolving conflicts at each level, as well as data extraction, data analyses, and data synthesis. COSMIN criteria include recommended study evaluation criteria for appraising quality and rigour of included psychometric studies. The current systematic review used the web-based collaborative SWIFT-Active Screener (Howard et al., 2020; Liu et al., 2023) review software for screening.

2.1. Search strategy

The original search was conducted on September 23, 2022, without any date restrictions. An updated search was conducted on May 30, 2023, without any date restrictions. The search used the following databases: PsycINFO, MEDLINE, APA PsycTests, Web of Science, ProQuest Theses & Dissertations, CINAHL, EMBASE, Health and Psychosocial Instruments. The following terms were used across databases: "moral* injury*", "moral* stress*", "moral* distress" (see Appendix A for string terms).

2.2. Inclusion and exclusion criteria

Inclusion criteria were: 1) studies that reported on the development and/or validation of a psychometric instrument designed to measure MD or MI; and 2) studies that implemented a measurement tool to quantitatively measure either MD or MI. Exclusion criteria were: 1) review studies or articles that provided a detailed plan of future study (e.g., review, meta-analysis, commentary, book, opinion piece, protocol papers); 2) studies that collected and analyzed only qualitative data (e.g., interview, focus groups); and, 3) studies not written in English or French.

2.3. Study selection

Ten screeners in total participated in the screening process. Articles were each independently screened by two raters at both the abstract and full-text level. Interrater reliability was good for title and abstract review (using percent agreement [85.9%], and Kappa [Fleiss and Conger; 0.718]) and full-text review (using percent agreement [95.8%], and Kappa [Fleiss and Conger; 0.915]). Conflicts were resolved at each screening level by study authors (S.H., N.E., & J.G.) until a consensus was reached. Included articles were then categorized as psychometric or measurement studies. Only psychometric studies were examined for the current review (see Fig. 1).

2.4. Data extraction and analyses

2.4.1. General

Articles were first assessed to identify and delineate the measurement model of each scale in support of conducting detailed quality evaluations of each psychometric study. Scales were identified as being

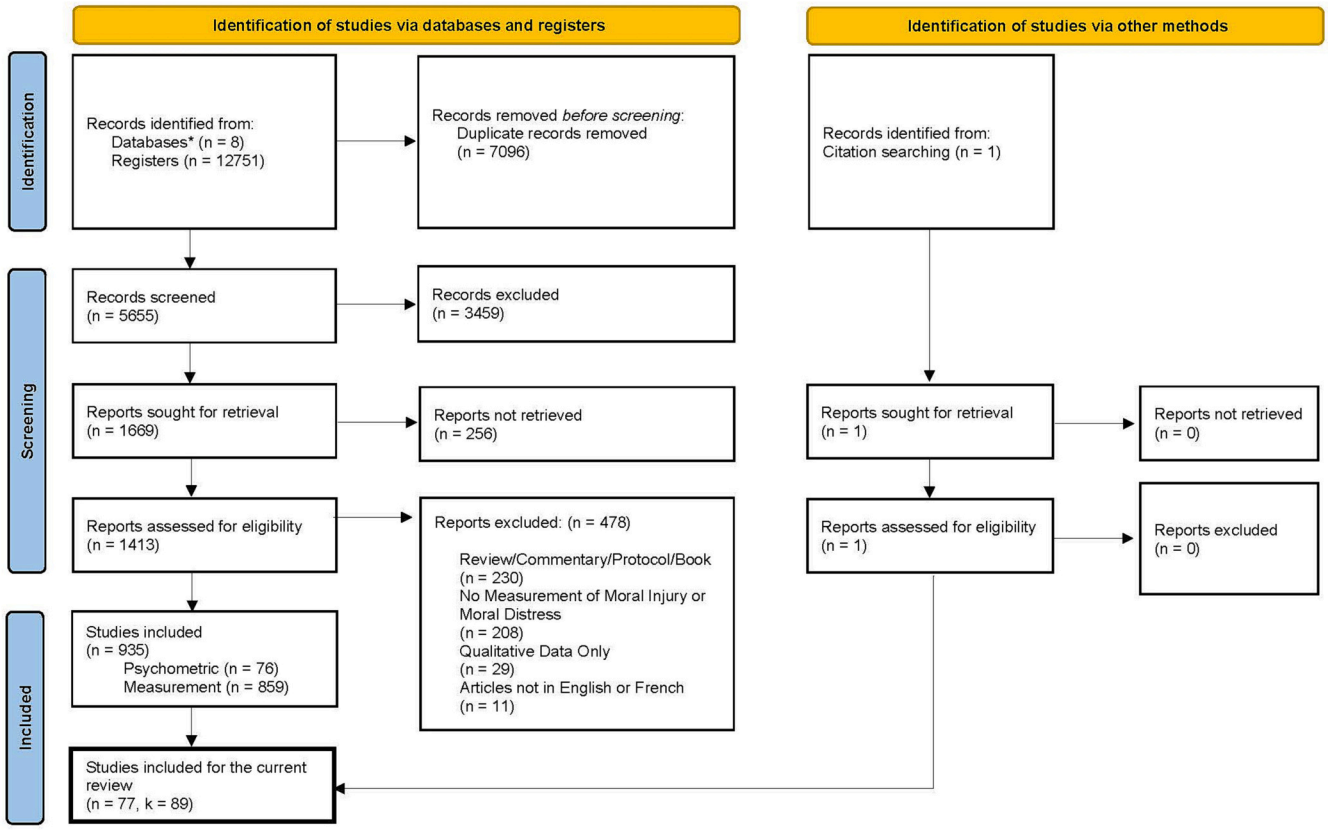

Fig. 1. Preferred reporting items for systematic reviews (PRISMA) flow diagram.

reflective, formative, or other. Scales categorized as "other" included items representative of both formative and reflective models (Coltman et al., 2008), and were evaluated as reflective, per COSMIN guidelines (Mokkink et al., 2018).

2.4.2. Systematic review

The following information was extracted from each article, as applicable (see below): 1) descriptive information (name of scale, type of scale model, optimal factor structure identified by authors, item composition of factors, population sampled, and sex information); 2) general design (definition of MD or MI and intended population for scale use); 3) formative model validity (whether the authors acknowledged the formative nature of the scale); 4) structural validity (e.g., factor analytic results, presence/absence of correlated residuals); 5) internal consistency (Cronbach's alpha or McDonald's Omega values); 6) test-retest reliability (e.g., test-retest conditions, correlations); 7) cross-cultural validity (e.g., scale translation details and psychometric comparability); 8) convergent or divergent validity evidence (e.g., correlations with related or unrelated scales). For articles reporting results from more than one study (i.e., scales were deployed in multiple separate groups or populations), the studies were extracted as independent. To support accuracy and consistency, there were two independent raters involved in data extraction (see Table S1 in Appendix A for a detailed description of each extracted variable).

Study evaluation criteria were based on COSMIN guidelines (see Mokkink et al., 2018) and adapted for the current review in consultation with psychometric and content experts. The quality of included articles was assessed using ratings across seven domains of reliability and validity, when applicable:

1. General Design: Was the construct of MD or MI clearly defined and grounded in a theoretical framework, and was the scale deployed in an appropriate population?
2. Formative Model Validity: Were formative measurement models appropriately acknowledged by authors?
3. Structural Validity: Were the psychometric tests appropriate for the type of scale, and were these tests conducted correctly?
4. Internal Consistency: How interrelated are the items on the scale? (e.g., Cronbach's alpha, McDonald's omega)?
5. Test-retest Reliability: Were test-retest conditions and statistical values appropriate?
6. Cross-cultural Validity: Were translation processes appropriate and described clearly?
7. Convergent and Divergent Validity: Did the convergent or divergent measures identified by authors correlate as would be expected with the MD or MI scales?

The domains evaluated for each study varied based on the type of model identified. In line with measurement principles for each model (Coltman et al., 2008), domains one, two, five, and six were evaluated for formative scales. For reflective scales and "other" scales, domains one and three to seven were evaluated to assess study quality. Results were presented with respect to ratings across domains (see Table S2 in Appendix A for detailed descriptions of each domain and the corresponding evaluation criteria).

2.4.3. Content analyses

Content analyses were performed on all available scales to specify the construct measured and explore thematic coverage for scales measuring MD or MI outcomes (i.e., including scales measuring both exposure and outcome, but excluding scales only measuring exposure; see Table S3 in Appendix A). All scale items were compiled into a single document and their order randomized. Each item was then independently coded by two coders on the following dimensions: 1) item type (i.e., exposure [e.g., witnessed a medication error] or outcome [e.g., I feel angry]); and 2) thematic content (for outcome items only; see Table S1 in Appendix A for details). Coding discrepancies were resolved through discussion and consensus, and items were then redistributed to their original scales. Item codes were then reviewed again in reference to scale instructions, with any corresponding code changes made thereafter reviewed by the first author and a research assistant, with any discrepancies resolved through discussion and consensus. Descriptive information about the instructions for each scale was also provided, including type of ratings and scoring, context/population specificity, indexed responses to a particular event, and time frame for response experiences (e.g., within the past month; see Table S1 in Appendix A).

3. Results

3.1. Study characteristics

The final sample consisted of 77 studies, incorporating 89 independent samples (refer to Appendix B for raw data and Appendix A for the references of included articles). Of these samples, 85 were sourced from peer-reviewed publications, while four were obtained from non-peer-reviewed sources, such as dissertations. A total of 42 unique scales were identified, all of which employed self-report measures to assess MD or MI. Within this group, 25 scales were labeled by their authors as MD assessments, while the remaining 17 were classified by their authors as MI assessments. Among the scales, 30 were found to measure both moral stressor exposures and associated outcomes (19 MD scales and 11 MI scales). Seven scales were solely focused on outcomes (comprising four MD scales and three MI scales) and three scales exclusively measured exposures (one MD scale and two MI scales). One MI scale was used to measure both exposure and outcomes in one study, while only measuring exposure in another study. Finally, one MD scale could not be definitively categorized due to a lack of access to the full scale and insufficient information provided in the scale development study (refer to Table S3 in Appendix A for detailed descriptive information).

Regarding measurement model, there were 21 scales identified as formative, 15 as reflective, four categorized as "other", one as mixed, and one deemed not applicable. The scale categorized as mixed was the Moral Injury Perpetration, Self-forgiveness, and Atonement Scales for Youth (MISY), which was considered formative in two studies but was categorized as "other" in another study based on methods used to evaluate the psychometric properties of the scale and final retention of items. The scale categorized as not applicable was the Moral Distress Thermometer (MDT), which consisted of only one item and, as such, could not be assigned a specific measurement model (see Table S3 in Appendix A).

There were 14 of the 42 scales included that were tested in more than one sample. Eight of these 14 scales had measurement models suitable for factor analysis (reflective or classified as "other"). Five of these eight scales were identified as having more than one optimal factor structure across samples. Studies exhibiting evidence of multiple factor structures were identified through one of two approaches: 1) scales showed different factor structures in different studies or samples (e.g., the Moral Injury Questionnaire – Military Version [MIQ-M]; two samples reported a 1-factor structure, and the other showed a 3-factor structure); or, 2) scales displayed similar factor structures across samples, but there was variability in the items retained (e.g., the Moral Injury Symptom Scale – Short Form [MISS-SF]; multiple samples displayed a 3-factor structure consisting of different items across samples). Scales exhibiting multiple factor structures were differentiated by assigning version numbers (e.g., V1, V2). The remaining three scales evidenced the same factor structure across samples, but all originated from the same source. For example, the Moral Outcomes of Relationship Aggression Scale (MORALS) consistently displayed the same 3-factor structure using samples reported in a single article (Taverna & Marshall, 2022; see Table S3 in Appendix A).

Scales varied regarding population context. Twenty-two scales were exclusively designed to measure MD within healthcare populations,

including physicians, nurses, medical students, and nursing students. Six scales were specifically designed to measure MI within military populations. Four scales assessed MI and two assessed MD in mixed populations (e.g., healthcare and military). Three MI scales and one MD scale were assessed only within the general population (i.e., civilians and/or undergraduate students). Four MI scales were only tested in other specific groups (refugees, journalists, and public safety personnel exclusively; see Table S3 in Appendix A).

Ten scales were tested in languages other than English; specifically Italian ($n = 4$), Persian ($n = 3$), Turkish ($n = 3$), German ($n = 2$), and Spanish ($n = 2$), as well as single instances of scales tested in Arabic, Chinese, Dutch, Farsi, Greek, Japanese, Swedish, and Tamil. Of these ten scales, six were tested in both English and another language, namely the MMD-HP, Moral Distress Questionnaire – de Veer (MDQ-dV); Moral Distress Questionnaire-Eizenberg (MDQ-E), MDS-R, Moral Distress Scale for Psychiatric Nurses (MDS—P), and MISS-SF. Further details can be found in Table S3 in Appendix A.

Sex was variably represented across study samples. There were 25 samples with >50% male participants, while 50 samples had fewer than 50% male participants. The percentage of males within a given sample was not provided for 14 samples (see Table S3 in Appendix A). Only 17 samples provided demographic data that included diverse gender identities. Additionally, 18 samples reported results from sex- or gender-based analyses, most of which were from healthcare populations ($k = 13$); however, many of these samples did not clearly differentiate between sex and gender in their analysis, an issue previously problematized (Callaghan, 2021). Information on sex- or gender-based analyses was extracted based on the information provided in the individual studies.

3.2. Systematic review

An analysis of the general design of all included samples ($k = 89$) revealed that the most frequent rating was '*very good*' ($k = 80$; 90%). The most common rating was '*adequate*' ($k = 8$; 42%) in samples assessing test-retest reliability ($k = 19$). The most common rating for cross-cultural validity in samples using translated scales ($k = 23$) was also '*very good*' ($k = 20$; 87%). For samples with formative scales ($k = 43$), formative model validity was predominantly rated as '*inadequate*' ($k = 31$; 72%). For samples using reflective or "other" scales ($k = 45$), structural validity was most frequently rated as '*doubtful*' ($k = 18$; 40%). Ratings of internal consistency were largely '*very good*' ($k = 18$; 40%). For convergent and divergent validity, the ratings were mostly '*doubtful*' ($k = 19$; 43%; refer to Table 1 for details).

Of the 14 scales that were tested with more than one sample, six were identified as formative ($k = 26$; i.e., MMD-HP, MDQ-dV, MDQ-E, Moral Distress Scale-Corley [MDS-C], MDS-R, MDS—P). For these six scales, most samples utilized to test formative scales were rated '*very good*' for general design ($k = 24$; 92%), but many demonstrated '*inadequate*' formative model validity ($k = 19$; 73%). Samples assessing test-retest reliability ($k = 6$) were primarily '*adequate*' ($k = 4$, 67%). Scales tested using languages other than English ($k = 15$) generally demonstrated '*very good*' cross-cultural validity ratings ($k = 14$; 93%). Study evaluation scores for these six formative scales were generally consistent across all relevant domains (see Table 1).

There were seven out of the 14 scales tested that had more than one sample and were identified as reflective or other (six reflective, one 'other'; $k = 32$); specifically, the Expression of Moral Injury Scale (EMIS-M), Expression of Moral Injury Scale-Short Form (EMIS-M-SF), MIES, Moral Injury Outcomes Scale (MIOS), MIQ-M, MISS-SF, and the MORALS. Most samples were rated '*very good*' for general design ($k = 29$; 91%). Samples assessing cross-cultural validity ($k = 4$) were all rated as '*very good*' ($k = 4$, 100%). The most common rating of structural validity was '*doubtful*' ($k = 13$; 41%), primarily associated with samples using the MISS-SF ($k = 5$) and MIES ($k = 4$). The majority of samples rated as '*very good*' or '*adequate*' in this category included those using the MIOS ($k = 5$), MIES ($k = 3$) and MIQ-M ($k = 2$). The most common rating of internal consistency was '*very good*' ($k = 13$; 41%), with most of the associated samples using the MORALS ($k = 3$), MIES ($k = 3$), and EMIS-M ($k = 3$). All samples rated as '*doubtful*' or '*inadequate*' were from studies using the MISS-SF ($k = 6$), MIES ($k = 5$), and MIQ-M ($k = 2$). Samples assessing test-retest reliability ($k = 5$) primarily showed '*inadequate*' ($k = 2$, 40%) or '*adequate*' ($k = 2$, 40%) ratings. For convergent and divergent validity ($k = 30$), the most common rating was '*doubtful*' ($k = 15$; 50%), with most of the associated samples using the MISS-SF ($k = 5$), EMIS-M ($k = 3$), EMIS-M-SF ($k = 2$), MIES ($k = 2$), and MIQ-M ($k = 2$). All samples with a '*very good*' rating of convergent and divergent validity came from studies of the MIOS ($k = 3$), MISS-SF ($k = 2$), and MORALS ($k = 2$), while most samples with '*inadequate*' ratings came from studies of the MIES ($k = 6$). The remaining scale, MISY ($k = 3$), was not evaluated due to its mixed model classification (i.e., two samples were classified as formative and one sample was classified as other; see Table 1).

The MORALS and MIOS were generally associated with higher ratings across multiple domains. Nevertheless, it should be noted that samples tested with the MIOS (Litz et al., 2022) and MORALS (Taverna & Marshall, 2022) were generated from a single publication. The EMIS-M and MIQ-M were associated with higher ratings in some domains and lower ratings in others. There were too few samples using the EMIS-M-SF ($k = 2$) to draw reliable conclusions. The MIES ($k = 6$) was associated with conflicting ratings within multiple domains (e.g., some '*very good*' or '*adequate*', others '*doubtful*'), but a larger proportion of samples had lower ratings within and across all domains. The MISS-SF was consistently associated with low ratings across all domains.

3.3. Study evaluation across constructs measured

Exposure. Among samples tested with scales exclusively assessing exposure to potential moral stressors ($k = 5$), the typical rating was '*very good*' for general design but '*inadequate*' for formative model validity. Samples assessing test-retest reliability ($k = 2$) were all rated as '*very good*' ($k = 2$; 100%). None of the scales that focused solely on exposure were tested in a language other than English (see Table 1).

Outcomes. Samples tested with scales exclusively measuring MD or MI outcomes ($k = 13$) were typically rated as having '*very good*' general design ($k = 11$; 85%). In terms of test-retest reliability, half of the four samples assessed were rated as '*inadequate*' ($k = 2$; 50%). Cross-cultural validity ratings were not applicable for all samples (i.e., scales were only available or tested in English). Samples tested with scales using reflective or other measurement models ($k = 12$) were typically rated as having '*adequate*' structural validity ($k = 5$; 42%) and '*very good*' internal consistency ($k = 6$; 50%). Convergent and divergent validity ratings ($k = 10$) were mostly '*very good*' ($k = 7$; 70%). There was one scale (MDT) consisting of only one item, thus it could not be classified as formative, reflective, or other, and consequently was not evaluated (see Table 1).

Exposure and Outcomes. Samples tested with scales measuring moral stressor exposures and outcomes ($k = 70$) were typically rated as having '*very good*' general design ($k = 64$: 91%). Samples assessing test-retest reliability ($k = 13$) were typically rated as '*adequate*' ($k = 8$; 62%). In evaluating cross-cultural validity (k = 23), most samples were rated as '*very good*' ($k = 20$; 87%). Among samples with formative scales ($k = 38$), most were rated as having '*inadequate*' validity ($k = 26$; 68%). Approximately half of samples tested using reflective and other scales ($k = 32$) were rated as having '*doubtful*' structural validity ($k = 15$; 47%), and less than half were rated as having '*very good*' internal consistency ($k = 12$; 38%). Convergent/divergent validity ratings were mostly '*doubtful*' ($k = 18$; 56%; see Table 1).

3.4. Study evaluation of convergent and divergent validity

There were 105 scales (comprising 310 independent correlations)

Table 1
Study evaluation scores across scales (k = 89).

	k	General Design (*k*)	*k*	Formative Model Validity (*k*)	*k*	Structural Validity (*k*)	*k*	Internal Consistency (*k*)	*k*	Test-Retest Reliability (*k*)	*k*	Cross Cultural Validity (*k*)	*k*	Convergent / Divergent Validity (*k*)	Average Study Rating[a] (*k*)
All Scales	89	✔✔✔ (80) ✔✔ (1) ✔ (7) × (1)	43	✔✔✔ (9) × (31) ■ (3)	45	✔✔✔ (4) ✔✔ (10) ✔ (18) × (13)	45	✔✔✔ (18) ✔✔ (11) ✔ (8) × (8)	45	✔✔ (2) ✔ (1) × (6) ■ (36)	89	✔✔✔ (20) ✔ (1) × (2) ■ (66)	44[+]	✔✔✔ (11) ✔ (19) × (14)	
AMIS[R]	1	✔✔✔ (1)			1	✔✔✔ (1)	1	✔✔✔ (1)	1	■ (1)	1	■ (1)	1	× (1)	2.25
BMIS-N[O]	1	✔✔✔ (1)			1	✔ (1)	1	✔✔ (1)	1	■ (1)	1	■ (1)	1	✔✔✔ (1)	2.25
BMIS-P[R]	1	✔✔✔ (1)			1	✔ (1)	1	✔✔ (1)	1	■ (1)	1	■ (1)	1	× (1)	1.50
BSMD-N[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	■ (1)			1.50
CCRS[R]	1	✔ (1)			1	× (1)	1	✔✔ (1)	1	■ (1)	1	■ (1)	1	× (1)	0.75
CES-M[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	■ (1)			1.50
COVID-MDS[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	■ (1)			1.50
EMIS-M[R]	3	✔✔✔ (3)			3	✔✔ (1) ✔ (1) × (1)	3	✔✔✔ (3)	3	✔✔ (1) ■ (2)	3	■ (3)	3	✔ (3)	2.00
EMIS-M-SF[R]	2	✔ (2)			2	✔ (2)	2	✔✔ (2)	2	■ (2)	2	■ (2)	2	✔ (2)	1.25
HWEDQ[F]	1	✔✔✔ (1)	1	× (1)					1	✔✔✔ (1)	1	■ (1)			2.00
INTEL-Values[R]	1	✔ (1)			1	✔ (1)	1	× (1)	1	✔✔✔ (1)	1	■ (1)	1	× (1)	1.00
IT-ESMEE[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	✔✔✔ (1)			2.00
MD-APPS[R]	1	✔✔✔ (1)			1	× (1)	1	✔✔ (1)	1	■ (1)	1	■ (1)	1	✔ (1)	1.50
MDDCS[O]	1	✔✔✔ (1)			1	× (1)	1	✔✔✔ (1)	1	■ (1)	1	■ (1)	1	× (1)	1.50
MDQ-A[F]	1	✔✔✔ (1)	1	✔✔✔ (1)					1	✔✔ (1)	1	■ (1)			2.67
MDQ-dV[F]	2	✔✔✔ (2)	2	× (2)					2	■ (2)	2	✔✔✔ (1) ■ (1)			1.80
MDQ-E[F]	2	✔✔✔ (2)	2	✔✔✔ (1) × (1)					2	✔✔ (1) ✔ (1)	2	✔✔✔ (1) ■ (1)			2.00
MDRS[F]	1	✔✔✔ (1)	1	✔✔✔ (1)					1	■ (1)	1	■ (1)			3.00
MDS-B[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	× (1)			1.00
MDS-C[F]	2	✔✔✔ (1) ✔ (1)	2	✔✔✔ (1) × (1)					2	× (1) ■ (1)	2	■ (2)			1.40
MDS-CN[F]	1	✔✔✔ (1)	1	✔✔✔ (1)					1	■ (1)	1	■ (1)			3.00
MDS-J[F]	1	✔✔✔ (1)	1	× (1)					1	✔✔ (1)	1	■ (1)			1.67
MDS-K[F]	1	✔✔✔ (1)	1	✔✔✔ (1)					1	■ (1)	1	✔ (1)			2.33
MDS-R[F]	12	✔✔✔ (12)	12	✔✔✔ (2) × (8) ■ (2)					12	✔✔ (2) ■ (10)	12	✔✔✔ (7) × (1) ■ (4)			2.09
MDS-P[F]	3	✔✔✔ (3)	3	× (3)					3	■ (3)	3	✔✔✔ (2) ■ (1)			1.88
MDSQ[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	■ (1)			1.50
MIAS[R]	1	✔✔ (1)			1	× (1)	1	× (1)	1	■ (1)	1	✔✔✔ (1)	1	× (1)	1.00
MDT*	1	✔✔✔ (1)									1	■ (1)	1	✔✔✔ (1)	3.00
MIDS[R]	1	✔✔✔ (1)			1	✔ (1)	1	✔✔✔ (1)	1	× (1)	1	■ (1)	1	✔✔✔ (1)	2.00
MIA-PSP[O]	1	✔✔✔ (1)			1	× (1)	1	✔✔✔ (1)	1	■ (1)	1	■ (1)	1	✔ (1)	1.75

(*continued on next page*)

Table 1 (*continued*)

	k	General Design (*k*)	*k*	Formative Model Validity (*k*)	*k*	Structural Validity (*k*)	*k*	Internal Consistency (*k*)	*k*	Test-Retest Reliability (*k*)	*k*	Cross Cultural Validity (*k*)	*k*	Convergent / Divergent Validity (*k*)	Average Study Rating[a] (*k*)
MIES[R]	8	✔✔✔ (7) ✔ (1)			8	✔✔✔ (1) ✔✔ (2) ✔ (4) × (1)	8	✔✔✔ (3) ✔ (3) × (2)	8	× (1) ■ (7)	8	■ (8)	8	✔ (2) × (6)	1.42
MIOS[R]	5	✔✔✔ (5)			5	✔✔✔ (1) ✔✔ (4)	5	✔✔✔ (2) ✔✔ (3)	5	■ (5)	5	■ (5)	3[+]	✔✔✔ (3)	2.61
MIQ-M[O]	3	✔✔✔ (3)			3	✔✔✔ (1) ✔✔ (1) × (1)	3	✔✔✔ (1) × (2)	3	■ (3)	3	■ (3)	3	✔ (2) × (1)	1.58
MISS-M[R]	1	✔✔✔ (1)			1	✔ (1)	1	✔✔✔ (1)	1	× (1)	1	■ (1)	1	✔ (1)	1.60
MISS-SF[R]	8	✔✔✔ (8)			8	✔✔ (1) ✔ (5) × (2)	8	✔✔✔ (1) ✔✔ (1) ✔ (5) × (1)	8	✔✔ (1) × (1) ■ (6)	8	✔✔✔ (4) ■ (4)	8	✔✔✔ (2) ✔ (5) × (1)	1.74
MISY**	3	✔✔✔ (3)	2	× (2)	1	× (1)	1	✔✔ (1)	3	✔✔✔ (1) ■ (2)	3	■ (3)	1	✔ (1)	1.67
MMD-HP[F]	5	✔✔✔ (4) × (1)	5	✔✔✔ (1) × (4)					5	✔✔ (1) ■ (4)	5	✔✔✔ (3) ■ (2)			1.86
MORALS[R]	3	✔✔✔ (3)			3	✔✔ (1) ✔ (1) × (1)	3	✔✔✔ (3)	3	✔ (1) ■ (2)	3	■ (3)	3	✔✔✔ (2) ✔ (1)	2.23
PIDS[R]	1	✔ (1)			1	× (1)	1	× (1)	1	× (1)	1	■ (1)	1	✔✔✔ (1)	0.80
TMIS-J[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	■ (1)			1.50
Unsp. MD-W[F]	1	✔✔✔ (1)	1	■ (1)					1	■ (1)	1	■ (1)			3.00
Unsp. MD-S[F]	1	✔✔✔ (1)	1	× (1)					1	■ (1)	1	■ (1)			1.50

Notes. ✔✔✔ = very good; ✔✔ = adequate; ✔ = doubtful, × = inadequate; ■ = not available (see Table S2 in Appendix A for corresponding evaluation criteria); *k* = total number of samples; (*k*) = number of samples for each rating; bolded numbers in 'All Scales' = most common rating; blank spaces indicate that a rating was not given based on the measurement model identified.

For *general design,* possible ratings included *inadequate*, *doubtful*, *adequate*, or *very good*. For *formative model validity,* possible ratings included *inadequate*, *very good* or *not available*. For *structural validity,* possible ratings included *inadequate*, *doubtful*, *adequate*, or *very good*. For *internal consistency,* possible ratings included *inadequate*, *doubtful*, *adequate*, or *very good*. For *test-retest reliability,* possible ratings included *inadequate*, *doubtful*, *adequate*, *very good*, or *not available*. For *cross-cultural validity,* possible ratings included *inadequate*, *doubtful*, *adequate*, *very good*, or *not available*. For *convergent/divergent validity,* possible ratings included *inadequate*, *doubtful*, *adequate*, or *very good*. *k* = number of independent samples; [F] = formative measurement model scale; [R] = reflective measurement model scale, [O] = other scales – these scales were treated as reflective for study evaluation; * = one item scale (formative and reflective scale ratings do not apply); ** = indicates that scale was classified as different measurement models across samples; [+] = indicates that a different sample was used for the analyses of convergent and divergent validity; [a] = average ratings were determined for each scale by summing the scores across domains, divided by the number of applicable samples. Scoring of study evaluation ratings: very good = 3, adequate = 2, doubtful = 1, inadequate = 0. Study evaluation criteria marked as N/A for a given study were not considered in the calculation; *Abbreviations.* AMIS = Adult Moral Injury Scale; BMIS-N = Brief Moral Injury Screen-Nieuwsma; BMIS-P = Brief Moral Injury Scale-Pfeffer; BSMD-N = Brazilian Scale of Moral Distress in Nurses; CCRS = C-Change Resident Survey – Moral Distress Subscale; CES-M = Combat Experiences Scale (Modified) Moral Injury and Atrocity Subscale; COVID-MDS = COVID-19 Moral Distress Scale; EMIS-M = Expression of Moral Injury Scale-Military Version; EMIS-M-SF = Expression of Moral Injury Scale-Military Version-Short Form; HWEDQ = Healthcare Workers Emergency Distress Questionnaire; INTEL-Values = Moral Distress Subscale of the Values of Intensive Care Nurses for End-of-Life; IT-ESMEE = Italian Moral Distress Scale for Nursing Students; MD-APPS = Moral Distress - Appraisal Scale; MDDCS = Moral Distress in Dementia Care Survey; MDQ-A = Moral Distress Questionnaire – Astbury; MDQ-dV = Moral Distress Questionnaire-de Veer; MDQ-E = Moral Distress Questionnaire-Eizenberg; MDRS = Moral Distress Risk Scale; MDS-B = Moral Distress Scale-Badolamenti; MDS-C = Moral Distress Scale-Corley; MDS-CN = Moral Distress Scale for Correctional Nurses; MDS-J = Moral Distress Scale-Jafari; MDS-K = Moral Distress Scale-Kleinknecht-Dolf; MDS-P = Moral Distress Scale for Psychiatric Nurses; MDS-R = Moral Distress Scale – Revised; MDSQ = Moral Distress Scale/Questionnaire; MDT = Moral Distress Thermometer; MIA-PSP = Moral Injury Assessment for Public Safety Personnel; MIAS = Moral Injury Appraisals Scale; MIDS = Moral Injury and Distress Scale; MIES = Moral Injury Events Scale; MIOS = Moral Injury Outcome Scale; MIQ-M = Moral Injury Questionnaire - Military Version; MISS-M = Moral Injury Symptom Scale – Military Version; MISS-SF = Moral Injury Symptom Scale – Military Version - Short Form; MISY = Moral Injury Perpetration, Self-forgiveness, and Atonement Scales for Youth; MMD-HP = Measure of Moral Distress for Healthcare Professionals; MORALS = Moral Outcomes of Relationship Aggression Scale, PIDS = Perpetration-Induced Distress Scale; TMIS-J = Toronto Moral Injury Scale for Journalists; Unsp. MD-W = Unspecified Moral Distress-Wiggleton; Unsp. MD-S = Unspecified Moral Distress-Sporrong.

used to measure either convergent or divergent validity (see Table 2). Each scale was classified under one of the following measurement categories (from most to least common): 1) mental health functioning and affect ($n = 20$); 2) work functioning ($n = 18$); 3) moral emotions ($n = 12$); 4) positive mental health ($n = 11$); 5) moral injury or distress ($n = 8$); 6) adverse exposure ($n = 7$); 7) pain and physical functioning ($n = 6$); 8) social functioning and support ($n = 6$); 9) other ($n = 5$); 10) religious and spiritual distress ($n = 4$); 11) alcohol and substance use ($n = 4$); and, 12) PTSD ($n = 4$; see Table S1 in Appendix A for details).

Across all scales used to measure either convergent or divergent validity ($n = 105$), 38 (36%) were used exclusively for convergent validity, 22 (21%) were used exclusively for divergent validity, and 26

Table 2
Reported convergent and discriminant correlations of moral injury / moral distress scales (n = 310).

Correlated Scale and Corresponding Category	*n*	Correlation			Outcome
		C	D	U	
Mental Health Functioning and Affect (k = 86)					
Maslach Burnout Inventory (Maslach & Jackson, 1981)	28	9	9	10	Burnout
Patient Health Questionnaire (Kroenke et al., 2001)	18	12	3	3	Depression
Depression Anxiety Stress Scale-21 (Lovibond & Lovibond, 1995)	8	2	–	6	Depression, Anxiety
Beck Depression Inventory (Beck et al., 1996)	4	2	–	2	Depression
Generalized Anxiety Disorder (Spitzer et al., 2006)	4	1	3	–	Anxiety
Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983)	4	4	–	–	Anxiety, Depression
Centre for Epidemiological Studies Depression Scale (Radloff, 1977)	3	2	–	1	Depression
Positive and Negative Affects Schedule (Watson et al., 1988)*	2	2	–	–	Negative Affect
State-Trait Anxiety Inventory for Children (Spielberger, 1973)	2	–	–	2	Anxiety
Suicidal Behaviors Questionnaire (Osman et al., 2001)	2	2	–	–	Suicidality
Professional Quality of Life Scale (Galiana et al., 2020)*	2	–	–	2	Compassion Fatigue, Burnout
Beck Anxiety Inventory (Beck et al., 1988)	1	1	–	–	Anxiety
Burnout Measure (Pines & Aronson, 1988)	1	–	–	1	Burnout
Compassion Fatigue-Short Scale (Dinç & Ekinc, 2019)	1	–	–	1	Compassion Fatigue
Big Five Personality Dimensions of Neuroticism (Rammstedt & John, 2007)	1	–	1	–	Negative Affect
General Health Questionnaire- 12-Item Short Form (Goldberg, 1972)	1	1	–	–	Mental Health
Kremen Anxiety Scale (Kremen, 1996)	1	–	–	1	Anxiety
Multidimensional Anxiety Questionnaire (Reynolds, 1999)	1	1	–	–	Anxiety
Multiscale Dissociation Inventory (Briere, 2002)	1	1	–	–	Mental Health
Second Victim Experience and Support Tool (Strametz, Siebold, Heistermann, Haller, & Bushuven, 2022)*	1	1	–	–	Mental Health
Work Functioning (k = 48)					
Copenhagen Psychosocial Questionnaire (Kristensen et al., 2005)	9	–	–	9	Ethical Climate/ Work Climate
General Nordic Questionnaire (Dallner et al., 2000)	9	6	3	–	Ethical Climate/ Work Climate
Second Victim Experiences and Support Tool (Strametz et al., 2022)*	6	6	–	–	Ethical Climate/ Work Climate
Turnover Intentions Questionnaire (Cohen, 1998)	5	3	–	2	Ethical Climate/ Work Climate
Ethical Environment Questionnaire (McDaniel, 1997)	4	3	–	1	Ethical Climate/ Work Climate
Potentially Mitigating Activities	2	2	–	–	Ethical Climate/ Work Climate
Quality Work Competence (Arnetz 1997; 1999)	2	–	–	2	Functioning at Work
Effect of Working During a Pandemic	1	1	–	–	Functioning at Work
Ethical Climate Subscale	1	–	1	–	Ethical Climate/ Work Climate
Hospital Ethical Climate Survey-Short Version (Olson, 1998)	1	–	–	1	Ethical Climate/ Work Climate
Index of Overall Job Satisfaction (Brayfield & Rothe, 1951)	1	–	–	1	Ethical Climate/ Work Climate
Intention to Quit	1	–	–	1	Functioning at Work
Job Satisfaction Scale	1	–	–	1	Functioning at Work
Olson's Hospital Ethical Climate Scale (Olson, 1998)	1	–	–	1	Ethical Climate/ Work Climate
The Work and Social Adjustment Scale (Mundt et al., 2002)	1	1	–	–	Functioning at Work
General Self-Efficacy Scale (Schwarzer & Jerusalem, 2010)	1	–	–	1	Ethical Climate/ Work Climate
Young-Schema Questionnaire Short Form (Young & Brown, 2005)	1	–	–	1	Functioning at Work
Professional Quality of Life Scale (Galiana et al., 2020)*	1	–	–	1	Functioning at Work
Moral Emotions (k = 44)					
Dimensions of Anger Reactions (Forbes et al., 2004)	9	4	–	5	Anger
State Shame and Guilt Scale (Marschall et al., 1994)	7	3	–	4	Guilt, Shame
Personal Feelings Questionnaire (Harder & Greenwald, 1999)	6	–	–	6	Guilt, Shame
Trauma-Related Guilt Inventory (Kubany et al., 1996)	5	5	–	–	Guilt
Guilt and Shame Proneness (Cohen et al., 2011)	4	–	–	4	Guilt, Shame
Other as Shamer Scale (Goss et al., 1994)	3	3	–	–	Shame
Adolescent Version of the Cook-Medley Hostility Index (Liehr et al., 2006)	2	–	–	2	Anger
Guilt Inventory-State Guilt Subscale (Marschall et al., 1994)	2	2	–	–	Guilt
Harder Personal Feelings Questionnaire (Harder et al., 1993)	2	–	–	2	Guilt, Shame
Trauma-Related Shame Inventory (Øktedalen et al., 2014)	2	2	–	–	Shame
Internalized Shame Scale (Cook, 1987)	1	1	–	–	Shame
Positive and Negative Affect Schedule-Guilt (Watson et al., 1988)*	1	1	–	–	Guilt
PTSD (k = 31)					
Post-Traumatic Stress Disorder Checklist (Weathers et al., 2013b)	25	17	1	7	PTSD
Brief Inventory of Psychosocial Functioning (Kleiman et al., 2020)	4	4	–	–	PTSD
International Trauma Questionnaire (Cloitre et al., 2018)	1	–	–	1	PTSD
Posttraumatic Cognitions Inventory (Foa et al., 1999)	1	–	1	–	PTSD
Moral Injury/Distress (k = 30)					
Expressions of Moral Injury Scale (Currier et al., 2017)	10	10	–	–	Moral Injury
Moral Injury Events Scale (Nash et al., 2013)	9	6	–	3	Moral Injury
Moral Distress Scale (Corley et al. 2001)	5	5	–	–	Moral Distress
Moral Injury Questionnaire (Currier et al., 2015)	2	1	–	1	Moral Injury
Moral Injury Symptoms Scale (Koenig et al., 2018)	1	–	–	1	Moral Injury
Measure of Moral Distress-Healthcare Professionals (Epstein et al., 2019)	1	1	–	–	Moral Distress
Moral Distress Thermometer (Wocial & Weaver, 2012)	1	1	–	–	Moral Distress
Self-Assessment of Moral Distress	1	1	–	–	Moral Distress

(continued on next page)

Table 2 (*continued*)

Correlated Scale and Corresponding Category	*n*	Correlation			Outcome
		C	D	U	
Religiosity/Spirituality (k = 18)					
Religious and Spiritual Struggles Scale (Exline et al., 2014)	13	8	–	5	Religiosity/ Spirituality
Religious Importance	2	–	2	–	Religiosity/ Spirituality
Spiritual Importance	2	–	2	–	Religiosity/ Spirituality
Belief into Action Scale (Koenig et al., 2014)	1	–	1	–	Religiosity/ Spirituality
Positive Mental Health (k = 16)					
Connor-Davidson Resilience Scale (Connor & Davidson, 2003)	3	3	–	–	Resilience
Warwick-Edinburgh Mental Wellbeing Scale (Tennant et al., 2007)	3	–	3	–	Positive Mental Health
Secure Flourish Index (VanderWeele, 2017)	2	–	2	–	Positive Mental Health
Scales of Psychological Wellbeing (Ryff, 1989)	1	–	–	1	Positive Mental Health
Self-Forgiveness	1	–	1	–	Positive Mental Health
Forgiving Others	1	–	1	–	Positive Mental Health
Adult Trait Hope Scale (Snyder et al., 1991)	1	–	1	–	Positive Mental Health
Gratitude Questionnaire (McCullough et al., 2002)	1	–	1	–	Positive Mental Health
Satisfaction with Life Scale (Diener et al., 1985)	1	1	–	–	Positive Mental Health
Positive and Negative Affect Schedule (Watson et al., 1988)*	1	1	–	–	Positive Mental Health
Second Victim Experience and Support Tool (Strametz et al., 2022)*	1	1	–	–	Resilience
Adverse Exposure (k = 9)					
Deployment Risk and Resilience Inventory (King et al., 2003)	2	–	–	2	Stressor/Trauma Exposure
Combat Experiences Scale (Guyker et al., 2013)	2	1	1	–	Stressor/Trauma Exposure
10-item Adverse Childhood Experiences Questionnaire (Felitti et al., 1998)	1	1	–	–	Adverse Childhood Experiences
Brief Warfare Exposure Scale (NASEM, 2018)	1	–	–	1	Stressor/Trauma Exposure
Impact of Event Scale-Revised (Horowitz et al., 1979)	1	1	–	–	Stressor/Trauma Exposure
Revised Conflict Tactics Scale (Straus et al., 1996)*	1	1	–	–	Stressor/Trauma Exposure
The Integration of Stressful Life Events Scale-Short Form (Holland et al., 2014)	1	1	–	–	Stressor/Trauma Exposure
Alcohol and Substance Use (k = 6)					
Alcohol Use Disorders Identification Test-Concise (Bush et al., 1998)	2	2	–	–	Alcohol and Substance Use
Alcohol Use Disorders Identification Test (Babor et al., 2001)	2	1	–	1	Alcohol and Substance Use
Alcohol Use	1	1	–	–	Alcohol and Substance Use
Drug Abuse Screening Test (Skinner, 1982)	1	–	–	1	Alcohol and Substance Use
Pain and Physical Functioning (k = 6)					
Difficulty Engaging in Physical Activity	1	–	1	–	Pain and Physical Functioning
Daily Physical Pain	1	–	1	–	Pain and Physical Functioning
Difficulty with Physical Activity	1	–	1	–	Pain and Physical Functioning
Severity of Daily Pain	1	–	1	–	Pain and Physical Functioning
Second Victim Experience and Support Tool (Strametz et al., 2022)*	1	1	–	–	Pain and Physical Functioning
Insomnia Severity Index (Morin et al., 2011)	1	1	–	–	Pain and Physical Functioning
Social Functioning and Support (k = 6)					
MSPSS (Zimet et al., 1988)	1	–	1	–	Social Functioning and Support
Social Involvement	1	–	1	–	Social Functioning and Support
Relationship Quality	1	–	1	–	Social Functioning and Support
Community Involvement	1	–	1	–	Social Functioning and Support
Support System	1	–	1	–	Social Functioning and Support
Propensity to Trust Scale (Frazier et al., 2013)	1	–	–	1	Social Functioning and Support
Other (k = 10)					
Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960)	3	2	–	1	Social Desirability
Years of Experience	3	2	–	1	Experience
Level of Education	2	–	2	–	Education
Revised Conflict Tactics Scale (Straus et al., 1996)*	1	1	–	–	Interpersonal Violence
World Health Organization Disability Assessments Schedule (WHO, 2010)	1	1	–	–	General Functioning

Note. Measures without citations were developed by the authors of the respective papers for use in their studies (see Appendix A for reference list of correlation measures used).

Abbreviations. C = convergent, D = divergent, U = unspecified; MMD-HP = Measure of Moral Distress for Healthcare Professionals; MSPSS = Multidimensional Scale of Perceived Social Support; NASEM = National Academies of Sciences, Engineering, and Medicine; WHODAS = World Health Organization Disability Assessments Schedule.

* indicates that the scale appears in multiple categories. In these cases, the authors used subscales of the named scale to capture various outcomes (e.g., Different items and subscales of the Positive and Negative Affect Schedule were used to measure positive affect, negative affect, and guilt).

(25%) were not explicitly identified as being used to assess convergent or divergent validity (i.e., unspecified). The remaining 19 (18%) scales were used inconsistently across samples: three scales were identified as either convergent, divergent, or unspecified, three scales were identified as either convergent or divergent, and 13 scales were identified as either convergent or unspecified (see Appendix A for a list of these scales).

There were 76 out of the 310 correlations that exclusively assessed convergent validity. The most common scales for examining convergent correlations specifically were the Expressions of Moral Injury Scale (Currier et al., 2017; n = 10; 13%), Second Victim Experiences and Support Tool (Strametz et al., 2022; n = 6; 8%), and the Trauma-Related Guilt Inventory (Kubany et al., 1996; n = 5; 7%). There were 28 correlations which assessed divergent validity exclusively. The most common scale for examining divergent correlations specifically was the Warwick-Edinburgh Mental Wellbeing Scale (Tennant et al., 2007; n = 3; 11%). Forty-eight correlations were exclusively unspecified. The most common scales for examining these unspecified correlations were Copenhagen Psychosocial Questionnaire (Kristensen, Hannerz, Høgh, & Borg, 2005; n = 9; 19%), Personal Feelings Questionnaire (Harder & Greenwald, 1999; n = 6; 13%), and Guilt and Shame Proneness (Cohen,

Wolf, Panter, & Insko, 2011; $n = 4$; 8%; see Appendix B).

The most commonly used scales across all correlations ($n = 310$) were the Maslach's Burnout Inventory (MBI; Maslach & Jackson, 1981; $n = 28$; 9%), the Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5; Weathers et al., 2013; $n = 25$; 8%), and the Patient Health Questionnaire (PHQ; Kroenke, Spitzer, & Williams, 2001; $n = 18$; 6%). The most commonly used scales for examining correlations specifically with MD ($n = 101$) were the MBI ($n = 21$; 21%), the Copenhagen Psychosocial Questionnaire (Kristensen et al., 2005; $n = 9$; 9%), and the General Nordic Questionnaire (Dallner et al., 2000; $n = 9$; 9%). The most commonly used scales for examining correlations specifically with MI ($n = 210$) were the PCL-5 ($n = 23$; 11%), the PHQ ($n = 15$; 7%), and Religious and Spiritual Struggles Scale (Exline, Pargament, Grubbs, & Yali, 2014; $n = 13$; 6%; see Appendix B).

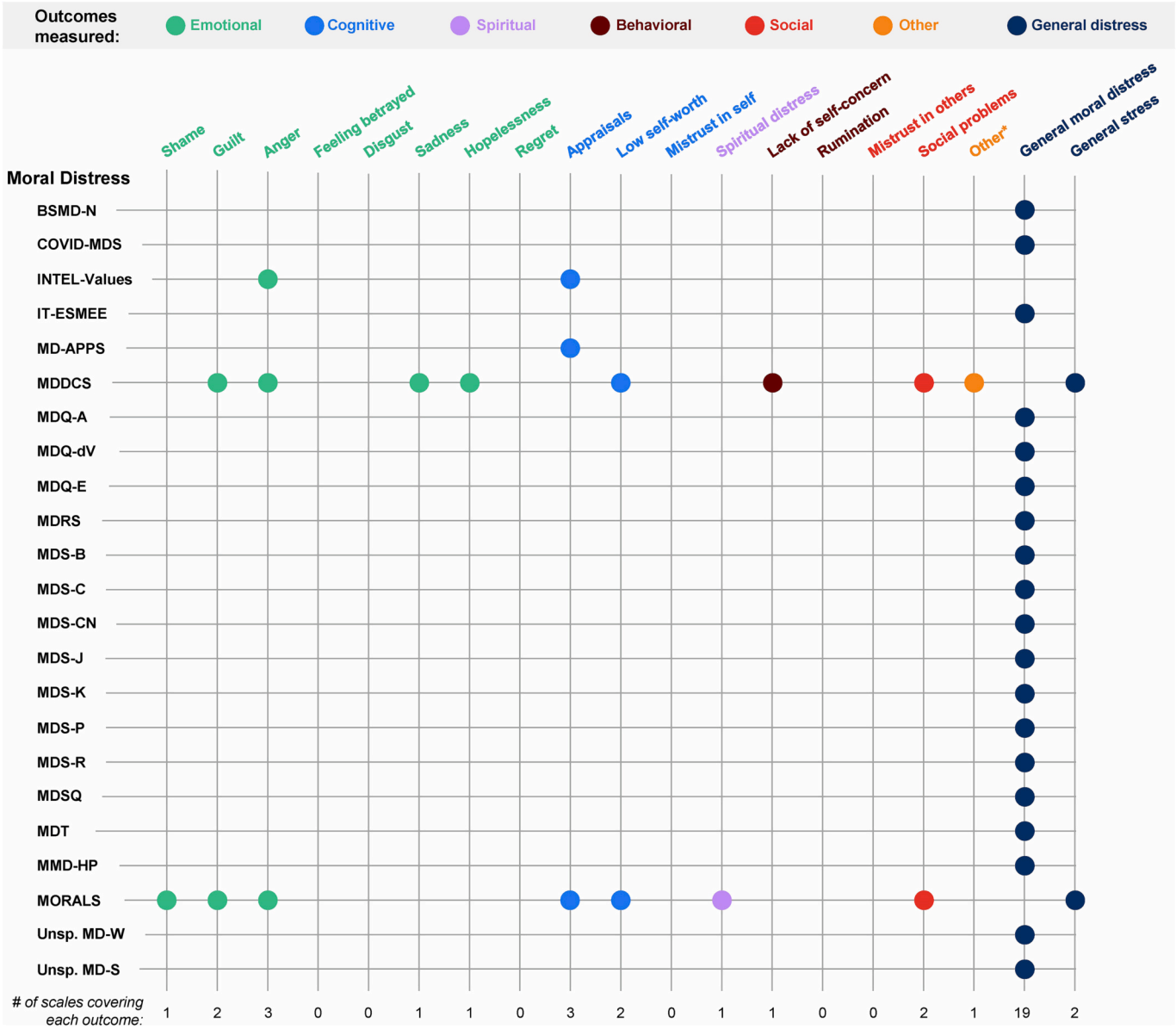

Fig. 2. *Content Themes Across Scales Measuring Moral Injury and Moral Distress Outcomes (Including Mixed Scales),* *Other = fear, physical consequences (e.g., sleep problems). *Abbreviations.* Moral Distress Scales: BSMD-N = Brazilian Scale of Moral Distress in Nurses; COVID-MDS = COVID-19 Moral Distress Scale; INTEL-Values = Moral distress sub scale of the Values of Intensive Care Nurses for End of Life; IT-ESMEE = Italian Moral Distress Scale for Nursing Students; MD-APPS = Moral Distress - Appraisal Scale; MDDCS = Moral Distress in Dementia Care Survey; MDQ-A = Moral Distress Questionnaire-Astbury; MDQ-dV = Moral Distress Questionnaire-de Veer; MDQ-E = Moral Distress Questionnaire-Eizenberg; MDRS = Moral Distress Risk Scale; MDS-B = Moral Distress Scale-Badolamenti; MDS-C = Moral Distress Scale-Corley; MDS-CN = Moral Distress Scale for Correctional Nurses; MDS-J = Moral Distress Scale-Jafari; MDS-K = Moral Distress Scale-Kleinknecht-Dolf; MDS-P = Moral Distress Scale for Psychiatric Nurses; MDS-R = Moral Distress Scale – Revised; MDSQ = Moral Distress Scale/Questionnaire; MDT = Moral Distress Thermometer; MMD-HP = Measure of Moral Distress for Healthcare Professionals; MORALS = Moral Outcomes of Relationship Aggression Scale; Unsp. MD-W = Unspecified Moral Distress-Wiggleton; Unsp. MD-S = Unspecified Moral Distress-Sporrong. Moral Injury Scales: AMIS = Adult Moral Injury Scale; BMIS-N = Brief Moral Injury Screen-Nieuwsma; BMIS-P = Brief Moral Injury Scale-Pfeffer; EMIS-M = Expression of Moral Injury Scale - Military; EMIS-M-SF = Expression of Moral Injury Scale-Short Form; MIA-PSP = Moral Injury Assessment for Public Safety Personnel; MIAS = Moral Injury Appraisals Scale; MIDS = Moral Injury and Distress Scale; MIES = Moral Injury Events Scale; MIOS = Moral Injury Outcome Scale; MIQ-M = Moral Injury Questionnaire - Military Version; MISS-M = Moral Injury Symptom Scale; MISS-SF = Moral Injury Symptom Scale -Short Form; MISY = Moral Injury Perpetration, Self-forgiveness, and Atonement Scales for Youth (Chaplo, 2015 version only); PIDS = Perpetration-Induced Distress Scale.

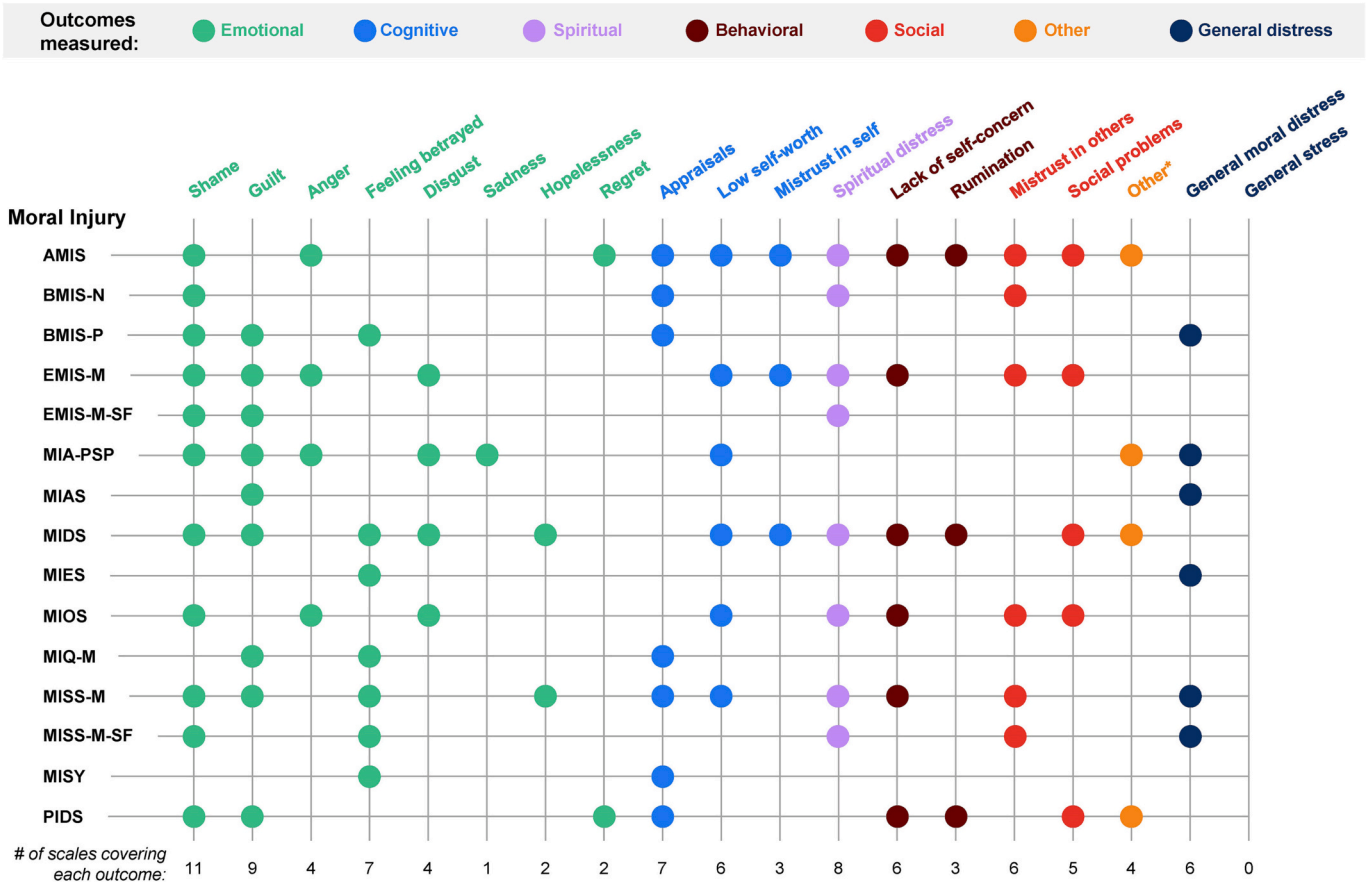

Fig. 2. (*continued*).

3.5. Content analyses

The content analyses included a total of 41 scales. One scale was unavailable for review (C-Change Resident Survey – Moral Distress Subscale [CCRS]). Descriptive information on the instructions for each scale is found in Table S4 (Appendix A), and details for thematic content coverage of scales measuring MD and MI outcomes are presented in Fig. 2a and b, respectively. Results demonstrate clear differences in scale content coverage between MD ($n = 23$) and MI ($n = 15$; see Fig. 2a and b, respectively). Scales putatively measuring MI included content spanning various outcome domains (i.e., emotional, behavioural, cognitive, spiritual, social, general distress, other), while MD scales tended to cover a single distress domain, either general MD ($n = 19$) or general stress ($n = 2$). The only MD scales that included content from multiple outcome domains were the Moral Distress Subscale of the Values of Intensive Care Nurses for End-of-Life (INTEL-Values), the Moral Distress in Dementia Care Survey (MDDCS), and the MORALS (see Fig. 2a). Uniquely, the Moral Distress - Appraisal Scale (MD-APPS) measures only the appraisal content domain, which is qualitatively different from all other unidimensional MD scales which only measure general MD. The MD scales assessing related outcome content ($n = 4$) included anger ($n = 3$), appraisals ($n = 3$), guilt ($n = 2$), and low self-worth/esteem ($n = 2$). The MI scales assessing related outcome content included shame ($n = 11$), guilt ($n = 9$), spiritual distress ($n = 8$), and appraisals ($n = 7$). No discernible qualitative differences were observed between scales of different measurement targets (outcomes only versus mixed scales).

Scale instructions typically directed respondents to consider a specific context ($n = 32$, one N/A; e.g., military experiences, work as a nurse) or allowed for broad contexts or populations ($n = 9$, with the exception of the MIES, which is context specific, except as adapted by Thomas, Bizumic, Cruwys, & Walsh, 2023). The majority of scales measuring outcomes ($n = 40$) applied a stressor-related framework to assess MD or MI ($n = 35$), either by asking participants to rate their distress in response to specific situations ($n = 17$) or by including elements of exposure and distress within items ($n = 6$; e.g., "I am troubled by morally wrong things I have done"). The remaining scales ($n = 12$) provided a stressor-related framework within scale instructions. Only three scales measuring outcomes paralleled the assessment of PTSD symptoms (e.g., as in the PCL-5) by asking respondents to keep a specific PMIE in mind while completing the scale (i.e., Moral Injury and Distress Scale [MIDS], MIOS, and MORALS). The MIDS and MIOS also provided a specified time frame for the outcomes (i.e., past month). Over half of scales ($n = 23$) provide no temporal reference point for item responses.

4. Discussion

The growing interest in moral stressors and their impacts underscores the critical need for reliable and valid measurement tools. We conducted a comprehensive review of current MD and MI scales, highlighting contemporary challenges that have substantial implications for researchers, clinicians, and organizations. The following discussion focuses on findings as they pertain to the operationalization, scale development and validation, and measurement for MD and MI, providing recommendations specific to each area.

4.1. Operationalization of MD and MI

Most scales of MD and MI were developed within the past decade, reflecting the relative novelty of this research area. While we found commonalities in the operationalization of MD and MI (i.e., distress stemming from events characterized by confrontations with one's and others' moral agency), there were also several important distinctions.

Scales using the term MD largely describe a general sense of distress specifically arising from exposure to pre-defined scenarios. In contrast, putative MI scales tended to prompt respondents to rate experiences across a diverse range of content domains, including specific emotions, beliefs, social impacts, and spiritual elements. Convergent validity for MD scales was commonly assessed using measures of work functioning and burnout, whereas convergent validity for MI scales was commonly assessed using measures of PTSD symptoms, major depressive disorder symptoms, or spiritual distress. These findings appropriately align with predominant conceptualizations of MD (Jameton, 1984) and MI (Litz et al., 2009), with MD chiefly construed as part of a broad network of experiences related to work functioning and burnout, and MI conceptualized as a mental and behavioural health problem. Notably, content analyses results indicated that certain MD scales (e.g., Taverna & Marshall, 2022) aligned with MI content domains. This finding may implicitly acknowledge that MD and MI lie on a continuum with shared features varying by intensity, underscoring the benefits to the field that would be provided by judicious delineations and operationalizations of these terms.

Content analyses results further suggest that scale developers generally concur on several fundamental outcome domains for MI, including moral emotions (e.g., shame, guilt) and spiritual distress; however, a high degree of variability across outcome domains remains. While challenges related to variability across outcome measures are not unique to MI (e.g., Fried, Flake, & Robinaugh, 2022), convergence towards a paradigmatic model that describes valid and reliable features of MI is likely to advance the field.

4.2. Recommendations

Consensus regarding MD and MI terminology will help support next steps in our understanding of moral stressors and their impacts. We therefore maintain that the terms “distress” and “injury” should be reserved for describing the *consequences* of exposure to potential moral stressors, acknowledging that such consequences are likely to exist along a continuum (Litz & Kerig, 2019). MI specifically should continue to be reserved for describing only those consequences which are potentially debilitating. Additional research is still needed to properly inform the potential boundary conditions of MI and its associated clinical utility. Similarities in content across extant MI scales, however, support preliminary consensus regarding core features. This agreement is particularly relevant as MD and MI research continues to expand beyond the commonly examined healthcare and military contexts. Importantly, any reporting of previous data should carefully consider the extent to which the scales applied reflect the intended operationalization of MD and MI. Moving forward, is is similarly important that researchers and clinicians make explicit the domains assessed by scales used to measure MD and MI in support of iterative improvements in this research area. This practice is particularly important when assessing the effectiveness of current and novel intervention approaches.

4.3. Scale development and validation

Most studies we reviewed appropriately conceptualized MD and MI as stressor-related phenomena. Still, the majority of formative scales were not acknowledged as such by their authors, who frequently applied reflective model psychometric approaches to scales assessing exposure to moral stressors. This error was especially evident among MD scales developed for the healthcare field, supporting concerns raised about the representation of MD in these contexts (Dean et al., 2020; Kolbe & de Melo-Martin, 2022). Disaggregating exposures from outcomes is likely to improve the utility of MD and MI scales and is a necessary next step in evaluating stressor-related models of these constructs.

A wide variety of scales were used to assess convergent and divergent validity, and we observed the inconsistent application of some scales across studies (e.g., the PCL-5 being used to assess convergent validity in some cases and divergent validity in others). Authors also frequently failed to specify hypothesized associations between a given scale and convergent or divergent indicators, which is problematic for construct validity. More robust assessments would involve a thorough evaluation of the psychometric quality of scales used to assess convergent or divergent validity (e.g., appropriately validated scales with good psychometric properties for the target population) and include correlational data across psychometric and measurement studies in aggregate (Mokkink et al., 2018). Such efforts would support validation of MD and MI, as well as inform clinical relevance and delineation from other outcomes (e.g., PTSD; Litz & Kerig, 2019).

4.4. Recommendations

Increased scrutiny concerning measurement models in the development and validation of MD and MI scales is sorely needed. As suggested by COSMIN (Mokkink et al., 2018), a simple “thought test” can guide researchers in this regard: If all items are expected to change when the construct changes (e.g., changes in MI would be reflected by changes in spiritual distress, guilt, *and* mistrust) the scale is reflective, and if not, the scale is formative. Scale developers can apply the same logic to each potential item within a scale.

Exposure scales need to be comprehensive in their content coverage. As noted recently by Karstoft and Armour (2023), variability in the specificity and range of content covered across scales risks misrepresenting population exposure prevalence and correlates. For example, a scale describing 10 emergency department PMIEs may demonstrate smaller associations with mental health outcomes than a scale including 30 PMIEs in that same context due to ceiling effects imposed by the shorter scale. In-depth consultation and a priori data collection on exposure types particular to a given setting will help enhance scale validity. Scales should also provide open-ended response options for describing PMIEs not assessed with existing items.

Scale instructions also warrant significant consideration, particularly regarding imagined responses. Asking participants to rate the *imagined* intensity with which they might respond to a given potential moral stressor ought to be avoided (e.g., Epstein et al., 2019). Including imagined responses artificially inflates scores, confounds the prevalence of real and imagined events, and obfuscates potential foci for intervention.

Substantial problems were noted with regard to sample size and factor analyses among studies assessing non-formative scales. Careful consideration of issues such as rotation method, correlation of residuals, appropriate thresholds for cross-loading and item retention, and interpretation of model fit indices should be conducted before data is collected (Osborne, 2008), all of which would help improve psychometric assessments of MD and MI scales. In assessing convergent and divergent validity, researchers need to include candidate scales exhibiting robust psychometric properties specifically for the intended population. Further, researchers need to take a clear a priori position on the expected direction, significance, and strength of correlated scales (Mokkink et al., 2018). Hypothesis testing regarding validity is required to properly delineate boundary conditions of the MD and MI constructs.

Regarding test-retest reliability, particular considerations for formative and reflective scales should be noted. Given the clinical relevance of MD and MI, it is important to properly understand the temporal stability of outcomes (measured using reflective models), as this has implications for assessment and outcome monitoring. In addition, MD and MI outcome assessments need to account for changes over time in the index stressor selected and time frame used for assessing outcomes. Appropriate assessment of test-retest reliability should include repeat administration at least two weeks after initial data collection using the same method and respondents (Mokkink et al., 2018). Notably, with respect to formative scales which, in this sample, all measured exposure, test-retest reliability should be interpreted with respect to the circumstances under which the scales are administered in

populations with an ongoing risk of exposure. For example, if formative scales are deployed among nurses working during an acute period of crisis, reported exposures to PMIEs may be more likely to change at retest. This would reduce the test-retest reliability estimate of the measure despite it accurately reflecting moral stressor exposure at both time points.

4.5. Measuring MD and MI

Most scales scored below adequate for structural validity based on the available sample results. Several of these ratings were from the MISS-SF and MIES (16% and 11% of ratings, respectively), which also produced different factor structures across studies. The current results may in part be due to the fact that moral stressor exposure and outcomes are often conflated by including items describing exposure (e.g., "I saw things that were morally wrong") and outcomes (e.g., "I am inclined to feel that I am a failure") or both (e.g., "I feel guilt over failing to save the life of someone in war"). The conflation of these features facilitates common method variance (Podsakoff, MacKenzie, & Podsakoff, 2012), such that observed variance may reflect event (exposure) commonalities rather than distress.

A similar problem was observed regarding the evaluation of internal consistency. While these ratings were generally good across studies, their validity is undermined because formative and reflective constructs were often conflated. For example, committing a medication error may be correlated with feelings such as guilt and shame in practice, but such a correlation would not reflect a latent construct (i.e., MI). Rather, the medication error is an antecedent event relative to a particular presentation of distress. Explicitly delineating exposures and outcomes when measuring MD or MI should substantially advance efforts to understand the mechanisms responsible for correlations between events and outcomes, which will ultimately inform the clinical and organizational utility of the MD and MI constructs.

Finally, the unequal distributions regarding sex we observed in this review were likely due to the populations commonly targeted for validation efforts; specifically, healthcare professionals were often recruited to complete MD scales (a population historically overrepresented by women), whereas military samples were used for MI scales (a population historically overrepresented by men). Few studies (22%) conducted sex- or gender-based analyses, and even fewer (19%) collected data on diverse gender identities. Additional research is needed to refine our understanding of the complex interactions between sex, gender, and mental health (Callaghan, 2021), and related measurement considerations. Additional research is also needed surrounding intersectional impacts associated with characteristics such as ethnicity, religion, function abilities, sexual orientation, and stressor exposures (Roberts, Austin, Corliss, Vandermorris, & Koenen, 2010; Roberts, Gilman, Breslau, Breslau, & Koenen, 2011), many of which are likely to be morally impactful in nature (e.g., witnessing a hate crime).

4.6. Recommendations

4.6.1. Non-formative scales

Recommendations are based on criteria adapted from COSMIN as well as guidelines proposed by Birnie, Hundert, Lalloo, Nguyen, and Stinson (2019) and Cohen et al. (2008) for reflective and other scales (see Table 3). Measurement of MD and MI is relatively novel compared to other developed areas of stressor-related constructs (e.g., PTSD). As such, existing guidelines were adapted to reflect the novelty of this area and to note that additional validation work is still required for most scales. Briefly, categories can be interpreted as follows: *Leading Recommendation* reflects scales which currently demonstrate consistently favourable conceptual and psychometric properties across multiple samples; *Provisional Recommendation* reflects scales which currently demonstrate adequate conceptual and psychometric properties in at least one sample; *Weak Recommendation Against* reflects scales with less

Table 3
Recommendation criteria for reflective scales.

Category	Criteria for recommendation	Scales identified by criteria
Leading Recommendation	• Evaluation in at least two samples published in peer-reviewed articles (see Table 1). • Scale appropriately reflects stressor-related construct framework (see Table 4). • Average study evaluation rating above adequate (> 2; see Table 2). • Combined average of ratings for structural validity, internal consistency and convergent/divergent validity above adequate (≥2; see Table 2). • Average ratings for structural validity, internal consistency and convergent/divergent validity all above adequate, respectively (≥ 2; see Table 2).	MIOS
Provisional Recommendation	• Evaluation in at least one sample published in a peer-reviewed article (see Table 1). • Scale appropriately reflects stressor-related construct framework (see Table 4). • Average study evaluation rating at least adequate (≥ 2; see Table 2). • Combined average of ratings for structural validity, internal consistency and convergent/divergent validity at least adequate (≥ 2). • No average rating for structural validity, internal consistency, or convergent/divergent validity below doubtful, respectively (< 1).	BMIS-N EMIS-M (Military) MIDS MORALS
Weak Recommendation Against	• Evaluation conducted in one or more samples published in a peer-reviewed article (see Table 1). • Average study evaluation rating below adequate (< 2; see Table 2). • Combined average of ratings for structural validity, internal consistency, and convergent/divergent validity below adequate (≤ 2).	BMIS-P CCRS EMIS-M (General population) EMIS-M-SF INTEL-Values MD-APPS MDDCS MIAS MIA-PSP MIES (HCWs) MISS-M MIQ-M PIDS
Supported Recommendation Against	• Evaluation conducted in multiple samples published in peer-review article(s) (see Table 1). • Average study evaluation rating below adequate (< 2; see Table 2). • Combined average of ratings for structural validity, internal consistency, and convergent/divergent validity combined below adequate (< 2). • Average ratings for structural validity, internal consistency, or convergent/divergent validity all below adequate, respectively (< 2).	MIES (Military) MISS-SF (Military, HCWs)
Undetermined	• Only one sample evaluated, not peer-reviewed (see Table 1); or • Only one sample in a peer-reviewed article but reporting only doubtful or N/A ratings for	AMIS EMIS-M (Public Safety Personnel) MIES (General population)

(continued on next page)

Table 3 (*continued*)

Category	Criteria for recommendation	Scales identified by criteria
	structural validity, internal consistency, and convergent/divergent validity (see Table 2)	MISY (Chaplo, 2015 version only)

Note. In cases where a scale is validated in multiple samples representing unique populations, criteria above were applied per population tested (see Table S3 in Appendix A). In cases where a proportion of samples is peer reviewed and the others are not, the non-peer reviewed samples were omitted in the relevant calculations for assessing recommendations.
Abbreviations. AMIS = Adult Moral Injury Scale; BMIS-N = Brief Moral Injury Screen-Nieuwsma; BMIS-P = Brief Moral Injury Scale-Pfeffer; CCRS = C-Change Resident Survey – Moral Distress Subscale; EMIS-M = Expression of Moral Injury Scale-Military Version; EMIS-M-SF = Expression of Moral Injury Scale-Military Version-Short Form; MD-APPS = Moral Distress - Appraisal Scale; MDDCS = Moral Distress in Dementia Care Survey; INTEL-Values = Moral Distress Subscale of the Values of Intensive Care Nurses for End-of-Life; MIA-PSP = Moral Injury Assessment for Public Safety Personnel; MIAS = Moral Injury Appraisals Scale; MIDS = Moral Injury and Distress Scale; MIES = Moral Injury Events Scale; MIOS = Moral Injury Outcome Scale; MIQ-M = Moral Injury Questionnaire - Military Version; MISS-M = Moral Injury Symptom Scale – Military Version; MISS-SF = Moral Injury Symptom Scale – Military Version - Short Form; MISY = Moral Injury Perpetration, Self-forgiveness, and Atonement Scales for Youth; MORALS = Moral Outcomes of Relationship Aggression Scale, PIDS = Perpetration-Induced Distress Scale.

than favourable conceptual and psychometric properties as evidenced in at least one sample; *Supported Recommendation Against* reflects scales demonstrating consistently inadequate psychometric properties across multiple samples; and *Undetermined* reflects scales which have not been peer-reviewed and/or for which only one peer-reviewed source is available which demonstrates low quality or unexamined psychometric properties.

Practical information relevant to at least provisionally recommended scales is summarized in Table 4. The MIOS was the only scale that met criteria for a "Leading Recommendation" for research and clinical use. The MIOS demonstrated consistently positive psychometric characteristics across multiple samples, uses broadly worded instructions, has broad content coverage, distinguishes exposures from outcomes, and indexes outcomes to a specific index event and a specified time frame (i.e., past month). The MIOS also screens for DSM-5 Criterion A exposures as well as MI-related functional impairments, supporting its clinical utility. Additional validation of the MIOS is, however, still required outside the military context.

All "Provisionally Recommended" scales appear appropriate for use in research and have the potential to be useful in public health and clinical contexts, with the MORALS, Brief Moral Injury Screen-Nieuwsma (BMIS-N), and EMIS-M currently limited by their context specificity.[1] The BMIS-N, in particular, has been designed as a screening instrument only for war-zone exposures and outcomes. The BMIS-N and MIDS include separate sections for assessing exposures and outcomes. The EMIS-M does not assess exposures, and outcomes are anchored to general military experiences. For the MORALS, outcomes are indexed to a specific behaviour measured using a separate scale (i.e., the Revised Conflict Tactics Scale, see Taverna & Marshall, 2022). Additional work is needed to assess sensitivity and specificity (e.g., clinical cut-points) for the MIOS and other "provisionally recommended" scales that may prove useful in clinical practice.

4.6.2. Formative scales

Moving forward, data from formative scales needs to be reported as descriptive information with item-level details. Reporting values for frequency and distress separately, where applicable, will be most informative. The clinical and organizational utility of MD requires delineating exposures and outcomes to facilitate proper identification of specific potential moral stressors and inform organizational intervention efforts (Kolbe & de Melo-Martin, 2022). The same is true for MI; however, the formative MI scales in the current review only assess PMIE exposures. Researchers applying formative scales should avoid inappropriately referring to sum or product scores as MD or MI outcomes, as the interaction between situational items and associated distress ratings can only serve to demonstrate the morally distressing nature of the situation assessed (i.e., exposure to moral stressors) rather than quantify distress as an outcome. Where there is interest in examining clinical outcomes pertinent to situations covered by formative MD and MI scales, we recommend using both a formative scale and a reflective scale or, where appropriate, a validated scale which includes a separate exposure section (e.g., the MIOS). This recommendation is consistent with assessments of potentially traumatic experiences (e.g., the Life Events Checklist; Weathers et al., 2013) and outcomes (e.g., the PCL-5; Weathers, Litz, et al., 2013). We also advise researchers and clinicians to consider the cultural applicability of selected scales. For example, scales assessing spiritual distress may include content pertinent to one or more specific religions, which may be inappropriate for individuals with diverse beliefs, including those who are neither religious or spiritual (Callaghan, 2023).

4.7. Limitations and future directions

Results and recommendations in the current review are based on relatively limited available psychometric work on MD and MI, constructs that continue to evolve. Replication and additional validation work are still needed to support the reliability and validity of existing MD and MI scales. Our review included studies describing only the development and psychometric validation of MD and MI scales; therefore, we are unable to comment on psychometric properties or conceptualization of MD and MI in measurement studies or other types of communications (e.g., editorials).

The COSMIN guidelines (Mokkink et al., 2018) were necessarily adapted for the current review to accommodate the relative novelty of MD and MI research. COSMIN does not provide guidance on the assessment of formative scales; however, most authors applied reflective psychometric assessment strategies to formative scales and criteria used to assess formative model validity were based on a liberal evaluation of methods. For example, principal components analysis (PCA) is a common dimension reduction method often conceptualized as appropriate for formative models and was considered as such for the current review. Recent critiques have argued that PCA relies on assumptions and techniques best represented by reflective measurement models and many recommend against applying PCA to formative scales (e.g., Mazziotta & Pareto, 2019). Accordingly, formative model validity results in the current review should be interpreted with caution. Scales classified as "other" were similarly evaluated for structural validity following COSMIN guidelines (i.e., were evaluated as reflective scales), but this was not entirely appropriate given the mixed nature of items in such scales, and scale developers should be attentive to creating scales that do not mix item types. In addition, the comprehensive set of COSMIN criteria for assessing scale design and development (e.g., item generation, comprehensiveness) were not applied in the current review due to resource limitations and high variability in design methods across studies. Content generated by systematic, bottom-up processes (e.g., targeted interviews, expert panel discussions) is likely to be most robust in supporting scales' general design and construct validity, and we suggest future evaluations and development of psychometric scales consider such processes.

The current review also did not evaluate the content or psychometric properties specific to any MD and MI subscales. Most scales were only

[1] The EMIS-M has been tested in other populations, but these samples did not demonstrate strong enough properties to be recommended at this time; see Table S3 in Appendix A.

Table 4
Practical information on using reflective scales that meet or exceed criteria for provisional recommendation.

	Leading Recommendation	Provisional Recommendations			
	MIOS	MORALS	BMIS-N	EMIS-M	MIDS
Validated Populations					
Military – General	✓	–	–	–	–
Military – Combat/War Zone	–	–	✓	✓	✓[a]
Healthcare Workers	–	–	–	–	✓[a]
First Responders	–	–	–	–	✓[a]
Perpetrators of IPV	–	✓	–	–	–
Context/Population Specificity[b]	Universal	IPV	Military – war zone	Military - General	Universal
Thematic Content Coverage					
Emotional	✓	✓	✓	✓	✓
Cognitive	✓	✓	✓	✓	✓
Spiritual	✓	✓	✓	✓	✓
Behavioural	✓	–	–	✓	✓
Social	✓	✓	✓	✓	✓
Language					
English	✓	✓	✓	✓	✓
Other	–	–	–	–	–
Separate Sections for Exposure & Outcomes	✓	–	✓	–	✓
Indexed Responses	✓	✓	–	–	✓
Specific Timeframe Provided	✓*	–	–	–	✓*
Number of Items[c]	14	15	4	17	18
Includes Subscales[d]	✓	–	–	✓	–
Appropriate Use					
Clinical - Screening	✓	✓	✓	✓	✓
Clinical – Outcome Monitoring[e]	✓	–	–	–	✓
Research	✓	✓	✓	✓	✓

Abbreviations: BMIS-N = Brief Moral Injury Screen-Nieuwsma; EMIS-M = Expression of Moral Injury Scale-Military Version; IPV = intimate partner violence; MIDS = Moral Injury and Distress Scale; MIOS = Moral Injury Outcome Scale; MORALS = Moral Outcomes of Relationship Aggression Scale.

[a] Grouped sample.

[b] As per scale instructions, scale prompts participants to reflect on a specific context (e.g., their military experiences; see Table 4.

[c] Excluding any exposure/indexing items.

[d] For outcome items only.

[e] Scales were deemed appropriate for clinical outcome monitoring if they provide a specific time frame for responses.

[*] Past month.

evaluated using a single sample and factor structures assessed with more than one sample often did not replicate. Indeed, efforts to delineate and describe self-related PMIE types (e.g., transgressing one's own values) compared with other-related PMIE types (e.g., being impacted by others' immoral behaviour) and their consequences has become central to MI theory and research (Griffin et al., 2020; Jordan, Eisen, Bolton, Nash, & Litz, 2017; Litz et al., 2018). Scales which appropriately delineate and evaluate these features of MI are likely to contribute substantially to developments in theory, research, and clinical applications. Future updates to the current review should consider examining the content and stability of subscales across time and contexts.

Furthermore, though certain members of our research team have previously contributed to the development and validation of MD and MI scales, recommendations criteria were applied systematically by members of our research team with no prior expertise in psychometric assessment of MD and MI. Together with the breadth of our research team, this serves to mitigate any potential bias in our recommendations. As previously mentioned, however, it is likely that as the field evolves future recommendations will be based on a more robust literature and are therefore subject to change.

Translation was the only aspect of cross-cultural validity assessed in the current review. Some scales (e.g., those using the term "G-d") may be restricted in their application, and formal assessment of cultural applicability is warranted. Occupational cultural differences can also differ substantially between countries and contexts (e.g., private vs. public healthcare systems; military objectives and ethos), and scales designed to assess context-specific exposures and outcomes may not replicate. Morality is substantively influenced by culture (Haidt, 2003), potentiating a wide variety of responses to contextually-dependent PMIEs. Additional research is needed to clarify the cross-cultural applicability of MD and MI scales.

Lastly, general scale design ratings in Table 1 were good for most studies, but the ratings only reflect whether scale developers appropriately described the theoretical foundations upon which their scales are based, and whether the scale was deployed in an appropriate population. General design goes beyond these features (Mokkink et al., 2018), and broader considerations of construct validity should be contextualized using all available information presented here (e.g., measurement model validity, whether a stressor-related framework was applied, and convergent or divergent validity).

5. Conclusion

Results show how the terms MD and MI are currently understood and applied in research. Several scales for assessing MD and MI were identified as appropriate for research and clinical use. There were substantial differences in the definitions and applications of the terms MD and MI across scales, signaling that a coherent paradigmatic understanding of the spectrum of MD and MI would help advance the field (e.g., Litz & Kerig, 2019). We identified several important limitations of the available literature, including inappropriate application of measurement models, and conflating of event exposures and outcomes. Preliminary research results support the clinical relevance of MI (Griffin et al., 2019); as such, ongoing coordination between scale developers, researchers, theorists, and clinicians is needed to enhance the development and application of appropriate mitigation and treatment strategies.

Role of funding sources

Funding for this study was provided by the Atlas Institute for Veterans and Families. Members of the funding organization were contributors to this work in the following capacities: Conceptualization,

project administration, reviewing and editing of the manuscript.

CRediT authorship contribution statement

Stephanie A. Houle: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Natalie Ein:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Julia Gervasio:** Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Rachel A. Plouffe:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Brett T. Litz:** Conceptualization, Methodology, Writing – review & editing. **R. Nicholas Carleton:** Conceptualization, Writing – review & editing. **Kevin T. Hansen:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing. **Jenny J.W. Liu:** Methodology, Writing – review & editing. **Andrea R. Ashbaugh:** Conceptualization, Methodology, Writing – review & editing. **Walter Callaghan:** Conceptualization, Writing – review & editing. **Megan M. Thompson:** Conceptualization, Writing – review & editing. **Bethany Easterbrook:** Conceptualization, Writing – review & editing. **Lorraine Smith-MacDonald:** Writing – review & editing. **Sara Rodrigues:** Writing – review & editing. **Stéphanie A.H. Bélanger:** Writing – review & editing. **Katherine Bright:** Writing – review & editing. **Ruth A. Lanius:** Writing – review & editing. **Clara Baker:** Data curation, Visualization. **William Younger:** Data curation, Formal analysis, Writing – review & editing. **Suzette Bremault-Phillips:** Conceptualization, Writing – review & editing. **Fardous Hosseiny:** Conceptualization, Resources, Writing – review & editing. **J. Don Richardson:** Conceptualization, Resources, Writing – review & editing. **Anthony Nazarov:** Conceptualization, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data for this review are available in Appendix B.

Acknowledgements

The authors would like to thank Michelle Birch for her assistance with this study. The Moral Injury Research Community of Practice includes Andrea Ashbaugh, Stéphanie Bélanger, Suzette Bremault-Phillips, Katherine Bright, Walter Callaghan, R. Nicholas Carleton, Bethany Easterbrook, Stephanie Houle, Fardous Hosseiny, Ruth Lanius, David Malloy, Margaret C. McKinnon, Anthony Nazarov, Sara Rodrigues, and Lorraine Smith-MacDonald, and Megan Thompson.

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cpr.2023.102377.

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425787 text rev.

Birnie, K. A., Hundert, A. S., Lalloo, C., Nguyen, C., & Stinson, J. N. (2019). Recommendations for selection of self-report pain intensity measures in children and adolescents: A systematic review and quality assessment of measurement properties. *PAIN, 160*(1), 5–18. https://doi.org/10.1097/j.pain.0000000000001377

Callaghan, W. (2021). Sex and gender: More than just demographic variables. *Journal of Military, Veteran and Family Health, 7*(s1), 37–45. https://doi.org/10.3138/jmvfh-2021-0027

Callaghan, W. (2023). Critical intercession for non-religious Canadian veterans on the intersections of moral injury, religion, and spirituality. *Journal of Military, Veteran and Family Health, 9*(2), 91–96. https://doi.org/10.3138/jmvfh-2022-0046

Cohen, L. L., La Greca, A. M., Blount, R. L., Kazak, A. E., Holmbeck, G. N., & Lemanek, K. L. (2008). Introduction to special issue: Evidence-based assessment in pediatric psychology. *Journal of Pediatric Psychology, 33*(9), 911–915. https://doi.org/10.1093/jpepsy/jsj115

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology, 100*(5), 947–966. https://doi.org/10.1037/a0022641

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61*(12), 1250–1262. https://doi.org/10.1016/j.jbusres.2008.01.013

Corley, M. C. (2002). Nurse moral distress: A proposed theory and research agenda. *Nursing Ethics, 9*(6), 636–650. https://doi.org/10.1191/0969733002ne557oa

Currier, J. M., Farnsworth, J. K., Drescher, K. D., McDermott, R. C., Sims, B. M., & Albright, D. L. (2017). Development and evaluation of the expressions of moral injury scale—Military version. *Clinical Psychology & Psychotherapy, 25*(3), 474–488. https://doi.org/10.1002/cpp.2170

Dallner, M., Elo, A.-L., Gamberale, F., Hottinen, V., Knardahl, S., Lindström, K., Skogstad, A., & Ørhede, E. (2000). *Validation of the general Nordic questionnaire (QPSNordic) for psychological and social factors at work*. Nordic Council of Ministers.

Dean, W., Talbot, S. G., & Caplan, A. (2020). Clarifying the language of clinician distress. *JAMA, 323*(10), 923–924. https://doi.org/10.1001/jama.2019.21576

Deschenes, S., Gagnon, M., Park, T., & Kunyk, D. (2020). Moral distress: A concept clarification. *Nursing Ethics, 27*(4), 1127–1146. https://doi.org/10.1177/0969733020909523

Easterbrook, B., Plouffe, R. A., Houle, S. A., Liu, A., McKinnon, M. C., Ashbaugh, A. R., … Nazarov, A. (2023). Moral injury associated with increased odds of past-year mental health disorders: A Canadian Armed Forces examination. *European Journal of Psychotraumatology, 14*(1), 2192622. https://doi.org/10.1080/20008066.2023.2192622

Epstein, E. G., & Hamric, A. B. (2009). Moral distress, moral residue, and the crescendo effect. *The Journal of Clinical Ethics, 20*(4), 330–342.

Epstein, E. G., Whitehead, P. B., Prompahakul, C., Thacker, L. R., & Hamric, A. B. (2019). Enhancing understanding of moral distress: The measure of moral distress for health care professionals. *AJOB Empirical Bioethics, 10*(2), 113–124. https://doi.org/10.1080/23294515.2019.1586008

Exline, J. J., Pargament, K. I., Grubbs, J. B., & Yali, A. M. (2014). The religious and spiritual struggles scale: Development and initial validation. *Psychology of Religion and Spirituality, 6*(3), 208–222. https://doi.org/10.1037/a0036465

Farnsworth, J. K., Drescher, K. D., Evans, W., & Walser, R. D. (2017). A functional approach to understanding and treating military-related moral injury. *Journal of Contextual Behavioral Science, 6*(4), 391–397. https://doi.org/10.1016/j.jcbs.2017.07.003

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology, 1*(6). https://doi.org/10.1038/s44159-022-00050-2. Article 6.

Giannetta, N., Villa, G., Pennestrì, F., Sala, R., Mordacci, R., & Manara, D. F. (2020). Instruments to assess moral distress among healthcare workers: A systematic review of measurement properties. *International Journal of Nursing Studies, 111*, Article 103767. https://doi.org/10.1016/j.ijnurstu.2020.103767

Griffin, B. J., Purcell, N., Burkman, K., Litz, B. T., Bryan, C. J., Schmitz, M., … Maguen, S. (2019). Moral injury: An integrative review. *Journal of Traumatic Stress, 32*(3), 350–362. https://doi.org/10.1002/jts.22362

Griffin, B. J., Williams, C. L., Shaler, L., Dees, R. F., Cowden, R. G., Bryan, C. J., … Maguen, S. (2020). Profiles of moral distress and associated outcomes among student veterans. *Psychological Trauma: Theory, Research, Practice, and Policy, 12*(7), 669–677. https://doi.org/10.1037/tra0000584

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford University Press.

Hall, N. A., Everson, A. T., Billingsley, M. R., & Miller, M. B. (2021). Moral injury, mental health and behavioural health outcomes: A systematic review of the literature. *Clinical Psychology & Psychotherapy, 29*(1), 92–110. https://doi.org/10.1002/cpp.2607

Hamric, A. B., Borchers, C. T., & Epstein, E. G. (2012). Development and testing of an instrument to measure moral distress in healthcare professionals. *AJOB Primary Research, 3*(2), 1–9. https://doi.org/10.1080/21507716.2011.652337

Harder, D. W., & Greenwald, D. F. (1999). Further validation of the shame and guilt scales of the Harder personal feelings Questionnaire-2. *Psychological Reports, 85*(1), 271–281. https://doi.org/10.2466/pr0.1999.85.1.271

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2022). Cochrane handbook for systematic reviews of interventions version 6.3. *Cochrane, 2022*. Available from www.training.cochrane.org/handbook.

Howard, B. E., Phillips, J., Tandon, A., Maharana, A., Elmore, R., Mav, D., … Shah, R. R. (2020). SWIFT-active screener: Accelerated document screening through active learning and integrated recall estimation. *Environment International, 138*, Article 105623. https://doi.org/10.1016/j.envint.2020.105623

Jameton, A. (1984). Nursing Practice: The Ethical Issues. Eweb:51336. https://repository.library.georgetown.edu/handle/10822/800986.

Jordan, A. H., Eisen, E., Bolton, E., Nash, W. P., & Litz, B. T. (2017). Distinguishing war-related PTSD resulting from perpetration- and betrayal-based morally injurious events. *Psychological Trauma: Theory, Research, Practice, and Policy, 9*(6), 627–634. https://doi.org/10.1037/tra0000249

Karstoft, K.-I., & Armour, C. (2023). What we talk about when we talk about trauma: Content overlap and heterogeneity in the assessment of trauma exposure. *Journal of Traumatic Stress, 36*(1), 71–82. https://doi.org/10.1002/jts.22880

Kolbe, L., & de Melo-Martin, I. (2022). Moral distress: What are we measuring? *The American Journal of Bioethics, 23*(4), 46–58. https://doi.org/10.1080/15265161.2022.2044544

Kristensen, T. S., Hannerz, H., Høgh, A., & Borg, V. (2005). The Copenhagen psychosocial questionnaire-a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian Journal of Work, Environment & Health, 31*(6), 438–449. https://doi.org/10.5271/sjweh.948

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kubany, E. S., Haynes, S. N., Abueg, F. R., Manke, F. P., Brennan, J. M., & Stahura, C. (1996). Development and validation of the trauma-related guilt inventory (TRGI). *Psychological Assessment, 8*(4), 428–444. https://doi.org/10.1037/1040-3590.8.4.428

Lamiani, G., Borghi, L., & Argentero, P. (2017). When healthcare professionals cannot do the right thing: A systematic review of moral distress and its correlates. *Journal of Health Psychology, 22*(1), 51–67. https://doi.org/10.1177/1359105315595120

Litz, B. T., Contractor, A. A., Rhodes, C., Dondanville, K. A., Jordan, A. H., Resick, P. A., … Peterson, A. L. (2018). Distinct trauma types in military service members seeking treatment for posttraumatic stress disorder. *Journal of Traumatic Stress, 31*(2), 286–295. https://doi.org/10.1002/jts.22276

Litz, B. T., & Kerig, P. K. (2019). Introduction to the special issue on moral injury: Conceptual challenges, methodological issues, and clinical applications. *Journal of Traumatic Stress, 32*(3), 341–349. https://doi.org/10.1002/jts.22405

Litz, B. T., Plouffe, R. A., Nazarov, A., Murphy, D., Phelps, A., Coady, A., … The Moral Injury Outcome Scale Consortium. (2022). Defining and assessing the syndrome of moral injury: Initial findings of the moral injury outcome scale consortium. *Frontiers in Psychiatry, 13*. https://doi.org/10.3389/fpsyt.2022.923928

Litz, B. T., Stein, N., Delaney, E., Lebowitz, L., Nash, W. P., Silva, C., & Maguen, S. (2009). Moral injury and moral repair in war veterans: A preliminary model and intervention strategy. *Clinical Psychology Review, 29*(8), 695–706. https://doi.org/10.1016/j.cpr.2009.07.003

Liu, J. W., Ein, N., Gervasio, J., Easterbrook, B., Nouri, M. S., Nazarov, A., & Richardson, J. D. (2023). Usability and accuracy of the SWIFT-ActiveScreener: Preliminary evaluation for use in clinical research. *medRxiv*. https://doi.org/10.1101/2023.08.24.23294573

Mantri, S., Lawson, J. M., Wang, Z., & Koenig, H. G. (2020). Identifying moral injury in healthcare professionals: The moral injury symptom scale-HP. *Journal of Religion and Health, 59*(5), 2323–2340.

Mantri, S., Lawson, J. M., Wang, Z., & Koenig, H. G. (2021). Prevalence and predictors of moral injury symptoms in health care professionals. *The Journal of Nervous and Mental Disease, 209*(3), 174–180.

Maslach, C., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior, 2*(2), 99–113. https://doi.org/10.1002/job.4030020205

Mazziotta, M., & Pareto, A. (2019). Use and misuse of PCA for measuring well-being. *Social Indicators Research, 142*(2), 451–476.

McCarthy, J., & Deady, R. (2008). Moral distress reconsidered. *Nursing Ethics, 15*(2), 254–262. https://doi.org/10.1177/0969733007086023

Mokkink, L. B., Prinsen, C., Patrick, D. L., Alonso, J., Bouter, L., de Vet, H. C., & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual, 78*(1), 6–63.

Morley, G., Ives, J., Bradbury-Jones, C., & Irvine, F. (2019). What is 'moral distress'? A narrative synthesis of the literature. *Nursing Ethics, 26*(3), 646–662. https://doi.org/10.1177/0969733017724354

Morris, D. J., Webb, E. L., Trundle, G., & Caetano, G. (2022). Moral injury in secure mental healthcare: Part I: Exploratory and confirmatory factor analysis of the moral injury events scale. *The Journal of Forensic Psychiatry & Psychology, 33*(5), 708–725. https://doi.org/10.1080/14789949.2022.2111318

Nash, W. P., Carper, T. L. M., Mills, M. A., Au, T., Goldsmith, A., & Litz, B. T. (2013). Psychometric evaluation of the moral injury events scale. *Military Medicine, 178*(6), 646–652. https://doi.org/10.7205/MILMED-D-13-00017

Nazarov, A., Fikretoglu, D., Liu, A., Thompson, M., & Zamorski, M. A. (2018). Greater prevalence of post-traumatic stress disorder and depression in deployed Canadian Armed Forces personnel at risk for moral injury. *Acta Psychiatrica Scandinavica, 137*(4), 342–354. https://doi.org/10.1111/acps.12866

Nillni, Y. I., Shayani, D. R., Finley, E., Copeland, L. A., Perkins, D. F., & Vogt, D. S. (2020). The impact of posttraumatic stress disorder and moral injury on women weterans' perinatal outcomes following separation from military service. *Journal of Traumatic Stress, 33*(3), 248–256. https://doi.org/10.1002/jts.22509

Oh, Y., & Gastmans, C. (2015). Moral distress experienced by nurses: A quantitative literature review. *Nursing Ethics, 22*(1), 15–31. https://doi.org/10.1177/0969733013502803

Osborne, J. W. (Ed.). (2008). *Best practices in quantitative methods*. Sage Publications Inc.

Plouffe, R. A., Easterbrook, B., Liu, A., McKinnon, M. C., Richardson, J. D., & Nazarov, A. (2021). Psychometric evaluation of the moral injury events scale in two Canadian armed forces samples. *Assessment, 30*(1), 111–123. https://doi.org/10.1177/10731911211044198

Plouffe, R. A., Nazarov, A., Forchuk, C. A., Gargala, D., Deda, E., Le, T., … Richardson, J. D. (2021). Impacts of morally distressing experiences on the mental health of Canadian health care workers during the COVID-19 pandemic. *European Journal of Psychotraumatology, 12*(1), 1984667. https://doi.org/10.1080/20008198.2021.1984667

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*(1), 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Riedel, P.-L., Kreh, A., Kulcar, V., Lieber, A., & Juen, B. (2022). A scoping review of moral stressors, moral distress and moral injury in healthcare workers during COVID-19. *International Journal of Environmental Research and Public Health, 19*(3), 1666. https://doi.org/10.3390/ijerph19031666

Ritchie, E. C. (2019). Reframing clinician distress: Moral injury not burnout. *Federal Practitioner, 36*(11), 506–507.

Roberts, A. L., Austin, S. B., Corliss, H. L., Vandermorris, A. K., & Koenen, K. C. (2010). Pervasive trauma exposure among us sexual orientation minority adults and risk of posttraumatic stress disorder. *American Journal of Public Health, 100*(12), 2433–2441. https://doi.org/10.2105/AJPH.2009.168971

Roberts, A. L., Gilman, S. E., Breslau, J., Breslau, N., & Koenen, K. C. (2011). Race/ethnic differences in exposure to traumatic events, development of post-traumatic stress disorder, and treatment-seeking for post-traumatic stress disorder in the United States. *Psychological Medicine, 41*(1), 71–83. https://doi.org/10.1017/S0033291710000401

Strametz, R., Siebold, B., Heistermann, P., Haller, S., & Bushuven, S. (2022). Validation of the German Version of the Second Victim Experience and Support Tool-Revised. *Journal of Patient Safety, 18*(3), 182–192.

Taverna, E., & Marshall, A. D. (2022). Development and validation of the moral outcomes of relationship aggression scale: A measure of moral distress following intimate partner violence perpetration. *Aggressive Behavior., 49*(1), 33–48. https://doi.org/10.1002/ab.22051

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., … Stewart-Brown, S. (2007). The Warwick– Edinburgh mental well-being scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes, 5*(1), Article 63. https://doi.org/10.1186/1477-7525-5-63

Thomas, V., Bizumic, B., Cruwys, T., & Walsh, E. (2023). Measuring civilian moral injury: Adaptation and validation of the Moral Injury Events Scale (Civilian) and Expressions of Moral Injury Scale (Civilian). *Psychological Trauma: Theory, Research, Practice, and Policy*.

Weathers, F. W., Blake, D. D., Schnurr, P. P., Kaloupek, D. G., Marx, B. P., & Keane, T. M. (2013). The Life Events Checklist for DSM-5 (LEC-5). Instrument available from the National Center for PTSD at www.ptsd.va.gov.

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). The PTSD checklist for DSM-5 (PCL-5). Scale available from the National Center for PTSD at www.ptsd.va.gov.

Zerach, G., Ben-Yehuda, A., & Levi-Belz, Y. (2023). *Prospective associations between psychological factors, potentially morally injurious events, and psychiatric symptoms among Israeli combatants: The roles of ethical leadership and ethical preparation.* Psychological Trauma: Theory Research Practice and Policy. https://doi.org/10.1037/tra0001466