Article:
Steen, S., Law, G.U. and Jones, C. (2025) The internal consistency of the moral injury event scale. European Journal of Psychological Assessment, 41 (6). ISSN: 1015-5759

https://doi.org/10.1027/1015-5759/a000824

Full Title: The Internal Consistency of the Moral Injury Event Scale: A Reliability Generalisation Meta-Analysis and Systematic Review

Data Availability: The data that support the findings of this review are openly available via the UK Data Service: Reference number(s): https://dx.doi.org/10.5255/UKDA-SN-856807; https://dx.doi.org/10.5255/UKDA-SN-856549.

Authors: Scott Steen, University of Hertfordshire; G. Urquhart Law (University of Birmingham); Chris Jones (University of Birmingham)

Abstract: The Moral Injury Event Scale (MIES) is a tool for measuring exposure to potentially morally injurious event(s) and distress. Although it reported acceptable psychometric properties in its initial development studies, it has since been used in multiple contexts and populations without assessment of its changing properties.. A reliability generalisation of the MIES and its Sub-Scales was therefore undertaken. A systematic search of electronic databases (PsychINFO; PTSD Pubs; MEDLINE; Scopus; Web of Science) identified 42 studies reporting internal consistencies (Cronbach's Alpha) up-to-April-2022. Unfortunately, few studies reported any other form of reliability or validity metric (e.g., test-retest, inter-rater reliability). A random effects model with a Bayesian analytic framework and the DerSimonian-Laird (1986) estimate was used. The review found the MIES to be an internally consistent tool based on alpha estimates at both Full-scale (α=.88; 95% CI: .87-.89) and Sub-scales (α=.82-.92; 95% CI: .79-.93). The review uncovered high heterogeneity and inconsistencies in its administration and modification although figures generally remained above acceptable levels (α≥.70). Based on the review, the MIES represents an internally

reliably tool for measuring potentially morally injurious events and distress at both Full and Sub-Scales according to pooled Cronbach's Alpha estimates.

Keywords: moral injury, potentially morally injurious event, meta-analysis, psychometrics, internal consistency

Corresponding author email address: s.steen@herts.ac.uk

ORCID: 0000-0002-6712-2761

Conflict of Interest: None declared.

Introduction

Background

In recent years, there has been significant progress in the field of Moral Injury (MI) research, particularly in measuring Potentially Morally Injurious Events (PMIEs) and MI-Symptoms (Koenig et al., 2019). MI is characterised by guilt, shame, loss of trust, and inner turmoil following a violation of deeply held moral beliefs (Litz et al., 2009; Shay, 1995). Conceptual models emphasise the impact of perceived transgressions and betrayal on an individual's moral code (Jinkerson, 2016; Litz et al., 2009; Shay, 1995). However, defining and assessing MI remains challenging due to the lack of consensus (Koenig et al., 2019). MI research has focused predominantly on military contexts, but recent work has expanded to include diverse settings and shared human experiences (McEwen et al., 2020). The mechanisms linking PMIEs to MI-Symptoms are debated, with cognitive models underlining event appraisal and negative attributions, and others proposing accumulated moral distress and complex-MI as possible factors (Farnsworth et al., 2017; Fleming, 2022; Nash, 2019).

Efforts to operationalise and measure MI have emerged in the last decade, including the development of dedicated psychometric tools (Koenig et al., 2019). These tools vary in scope, focusing on PMIEs, MI-Symptoms, and specific populations, with many designed for military contexts (Koenig et al., 2019). Tools developed within military contexts that are multidimensional (assessing PMIEs and MI-Symptoms) include the Moral Injury Event Scale (MIES) (Nash et al., 2013), Moral Injury Scale (Williamson et al., 2020), Moral Injury Outcome Scale (Yeterian et al., 2019), and Modified Military Moral Injury Questionnaire (Hodgson et al., 2021). The Moral Injury Symptoms Scale-Military Version (Koenig et al., 2018) represents a single-dimensional (MI-Symptoms) tool developed within a military context. There are also tools designed for non-military contexts, such as the Moral Injury

Appraisals Scale (Hoffman & Nickerson, 2021), Moral Injury Exposure and Symptom Scale-Civilian (Fani et al., 2021), and Moral Injury Scales for Youth (Chaplo et al., 2019).

The MIES is one of the most commonly used assessment tools in the field (Koenig et al., 2019) and has been integrated into other metrics (Chaplo et al., 2019; Fani et al., 2021; Hodgson et al., 2021; Hoffman & Nickerson, 2021; Koenig et al., 2018). While initially developed for military samples, adaptations have been made for non-military contexts (e.g., Animal Shelter Employees, Andrukonis & Protopopova, 2020; Parents involved with Child Protection Services, Haight et al., 2017; General Population, Khan et al., 2021). The initial design study included two Sub-Scales of Transgression (6-items) and Betrayal (3-items) (Nash et al., 2013), while a subsequent follow-up split the Transgression Sub-Scale into Transgression-Other (2-items, e.g., '*I saw things that were morally wrong*') and Transgression-Self (4-items, e.g., '*I acted in ways that violated my own moral code*'), while retaining the Betrayal category (3-items, e.g., '*I feel betrayed by leaders who I once trusted*') (Bryan et al., 2016). The MIES items lack temporal features, so ratings reflect generalised and ongoing experiences (Nash et al., 2013). The items are rated from 1 ('*Strongly Agree'*) to 6 ('*Strongly Disagree'*) with lower scores indicating higher PMIEs and MI distress; however, there are examples where the scoring is reversed. The tool does not include any clinical thresholds or severity bandings. Several authors have arbitrarily applied score thresholds to indicate endorsement of scales with most using above 3 ('*Slightly to Strongly Agree*') (Haight et al., 2017; Maguen et al., 2020a; Maguen et al., 2021; Sugrue, 2020). The initial design studies reported good internal consistencies at Full-Scale (α=.90) and Sub-Scales (α=.82-.89) and although these findings were promising, both recommended further evaluation (Bryan et al., 2016; Nash et al., 2013).

Despite the growing interest in MI assessment, there is a lack of consensus and limited validity and reliability testing for most measures (Hodgson et al., 2021; Williamson et al., 2020; Yeterian et al., 2019). The diverse nature of MI assessment is evident, not only within the MIES but across the field (Koenig et al., 2019). Given the limited synthesis of psychometric tools within the field, an improved understanding would help stakeholders in their choice of appropriate measures. As the earliest and most commonly used tool, the MIES represents a valuable candidate for synthesising and evaluating the MI assessment field.

Objectives

Systematically pooling psychometric properties using meta-analyses helps generate robust estimates about a tool's qualities and the contexts and characteristics influencing its reliability, limiting the erroneous practice of assuming transferable properties from different administrative contexts. MI assessment relies on suitable measures that have demonstrated validity and reliability. Unfortunately, the systematic search revealed little to no validity metrics meaning only the reliability property of internal consistency was analysed. Likewise, due to inconsistencies in reporting, it was not feasible to develop collective average MIES scores. For this review, a meta-analysis and systematic review assessed the MIES' psychometric properties, asking:

1. What are the MIES' collective internal consistencies?
2. How are internal consistencies influenced by different sample characteristics (e.g., age, gender), study design (e.g., location, correlational), or assessment method (e.g., payment, item modification)?

The review follows the REGEMA checklist, a widely recognised guideline for assessing the methodological quality and reporting standards of meta-analyses for reliability generalisation (Sanchez-Meca et al., 2021).

Methodology

Selection Criteria

Table 1 reports the selection criteria and their rationale. Selected studies included those using the MIES in any capacity, including wording adaptations, and reporting reliability and validity data (Cronbach's Alpha, Kappa, Intraclass Correlation Coefficient, Spearman/Pearson's r), either at Full-Scale or Sub-Scales. There were no restrictions on the publication date and target populations. Only publications written in English were eligible and translations of the MIES were accepted with English-written manuscripts. Original, empirical, and peer-reviewed studies using the measure in any capacity, including item changes, were selected. Secondary findings (e.g., systematic reviews), qualitative studies, sample sizes below 10, discussion, theoretical or position papers, book chapters, conference proceedings, and dissertations, were excluded.

Table 1

Selection Criteria and Accompanying Rationale

Search Strategies

A systematic search of studies reporting the MIES' reliability and validity was undertaken between June-2021-to-April-2022 via electronic databases (PsychINFO; PTSD Pubs; MEDLINE; Scopus; Web of Science). Boolean search terms and MeSH headings captured Moral Injury (MORAL, MORAL INJUR*, MORALLY INJURIOUS, TRANSGRESS*,

BETRAY*) and the MIES (MORAL INJURY EVENT* SCALE) along with articles citing the original design studies (Bryan et al., 2016; Nash et al., 2013). A Google Scholar alert for "MORAL INJURY EVENT* SCALE" was also set-up between June-2021-to-April-2022. There were no date restrictions on the searches.

Data Extraction

The corresponding author (SS) screened records by initially reviewing titles and abstracts, followed by a full-text review. Due to practical and resource constraints and based on the results from the subset, a random proportion of the 162 papers at the full-text review stage (10%, k=16) were independently cross-validated for eligibility by corresponding authors and were consistent in all inclusion decisions. All data were extracted by the author (SS) which included the internal consistencies of the MIES as the only psychometric property reported on. Study (Psychometric Design; Cross-sectional) and sample characteristics (Population (e.g., Military; Police); Location (e.g., US); Setting (e.g., Clinic; Community); Age (years); Gender (Male (%)); Ethnicity (White/Caucasian (%)); Education (College/University (%); Marital Status (Currently Married (%); Army Military Branch (%); Assessment Method (Interview; Online; Mixed); Payment (Paid; Not Paid; Unclear) were extracted. Sample characteristics were chosen based on the variable uses of the MIES and to gauge their effects on internal consistencies through moderator and meta-regression analyses. Although the initial searches aimed to generate score averages for the Full-Scale and Sub-Scales, this was not possible due to the many inconsistencies in scoring, calculating, and reporting across records.

Reported Reliability

While there are several methods for evaluating reliability and validity, an initial search revealed most sources solely reported the MIES' internal consistency as Cronbach's Alpha (α) (Cronbach & Shavelson, 2004). All meta-analytic syntheses were conducted on the raw alpha

coefficients. These data reflect item or Sub-Scale correlations, with generally acceptable levels rated minimally above α=.70 (Nunnally, 1975) to ideally above α=.90 (Nunnally & Bernstein, 1993). For this review, the null effect was therefore set to α=.70. Consistent with other research and guidelines (Cicchetti, 1994; Ponterotto & Ruckdeschel, 2007), pooled alpha estimates were classified '*Excellent*' (α>.89), '*Good'* (α=.85–.89), '*Moderate'* (α=.80–.84), '*Fair'* (α=.75–.79), or '*Unsatisfactory'* (α<.75).

Estimating the Reliability Induction and Other Sources of Bias

A set of quality criteria assessed the risk of bias within the selected papers by adapting relevant frameworks. As there are no standardised guidelines for psychometric properties of non-diagnostically based constructs like MI, the Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) (Whiting et al., 2011) criteria informed six categories of bias including Selection, Performance, Reporting, Detection, Statistical, and Generalisability. The QUADAS-2 is considered the research standard for assessing the quality of studies validating diagnostic tests (Venazzi et al., 2018) and represents established, comprehensive, and transparent criteria for bias ratings. Papers were scored Low (2), Unclear (1), or High (0) in each category based on the quality criteria (Steen, 2023a, Supplementary Table 1).

The corresponding author (SS) rated studies according to their designs including whether they specifically assessed the MIES psychometric properties or whether the tool was used in cross-sectional correlational designs for researching other primary aims. Psychometric design studies scored higher (30) than cross-sectional designs (20) to differentiate these characteristics and exceed the maximum risk of bias scores (12), thereby reflecting the review's aims to prioritise the assessment of psychometric properties. Psychometric design studies focus

specifically on evaluating the measurement properties and therefore involve comprehensive procedures for developing, refining and validating tools. These studies invest significant effort into collecting data to support the interpretation and use of the tools whereas cross-sectional designs have different priorities and typically do not have the same emphasis on examining the psychometric properties.

Data Extraction of Inducing Studies

As recommended by the REGEMA checklist (Sanchez-Meca et al., 2021), reliability induction was assessed based on estimation, induction, and omission. The reliability induction rate was calculated as the rate of studies reporting estimations (based on the current sample) against those inducting (referring to reliability estimations from other studies) or omitting (no reference to reliability). To gauge possible overestimations of internal consistencies, the estimation and inducing and omitting studies were compared using t-tests for continuous variables, including MIES test scores (Means (M); Standard Deviations (SD)), Age (M; SD), Gender (% Male), and Ethnicity (% White/Caucasian). Due to the inconsistencies in calculating and reporting MIES (M; SD) scores, those using similar calculations were compared.

Reliability of Data Extraction

The corresponding author (SS) weighted each criterion within the risk category equally during scoring. Due to practical and resource constraints and based on the results from the subset, a random sample of the 42 selected studies (10%; k=4) was independently cross-validated for risk of bias by corresponding authors and any disagreements were resolved by consensus. Only one paper required resolution by consensus relating to the Detection and Performance Bias domains, namely whether the study included probable word changes to the MIES and whether participants were asked to speak about their responses. The inter-rater reliability was therefore 92% for individual bias ratings and 75% for the consensus of the sample studies.

Transformation Method

The alpha coefficients for different samples within studies were not transformed or averaged as all estimates were extracted according to the reported samples. For example, Bryan et al. (2016) Samples 1 and 2 are included as individual sources, with the estimates extracted directly from the source and without transformation.

Statistical Model and Weighting Method

Given the variation in participant characteristics, study design, and test implementation, the random effects model was employed for all meta-analytic syntheses. The random effects model is suitable when significant variations exist between studies, offering greater robustness compared to the fixed effects model. In the random effects model, the calculation involved weighting by inverse variance with a Bayesian analytic framework and the DerSimonian-Laird (1986) estimate to quantify between studies variation and stabilise and manage variability. Although the estimator has been critiqued for its bias towards moderate-to-large heterogeneity

(Langan et al., 2018), it is widely used in medical and clinical research (DerSimonian & Laird, 2015) and has been shown to perform well in terms of bias and mean squared error (Sidik & Jonkman, 2006). It is available in most software packages (Wang et al., 2023), is less complex and non-iterative making it implementable (Chen et al., 2012; Makambi & Lu, 2013), and is appropriate given the number of studies included (Bender et al., 2018; Guolo & Varin, 2015). The distribution of primary study effects indicated little evidence of non-normality in the distribution of alpha coefficients across all levels using the DerSimonian-Laird (1986) estimator for between-studies variance (tau^2), supporting the use of the random-effects model. The omnibus tests for the Full-Scale and Sub-Scale reported significant heterogeneity ($p<.001$).

Heterogeneity Assessment and Moderator Analyses

If a study level alpha coefficient displays variation in the meta-analytic synthesis that cannot be accounted for by differences in sample size, it is considered heterogeneous. Heterogeneity can arise from methodological discrepancies, measurement errors, or uncontrolled individual differences within the body of literature. Higgins I^2 is a commonly used measure, where higher values indicate variation in effect that cannot be attributed to true variation in the population's effect distribution (Higgins et al., 2003). Considering the substantial methodological variation among the primary studies, heterogeneity was assessed using the I^2 statistic (>50%) and Cochran's Q statistic ($p>.10$) and considerable heterogeneity (which may complicate interpretation) was defined as $I^2>75\%$ (Higgins et al., 2003). When such high-level heterogeneity was observed, subsequent analyses focused on identifying the sources of heterogeneity between the estimates of internal reliability. This is accomplished through subgroup analyses (for categorical moderators; e.g., gender, ethnic origins) and meta-regression (for continuous moderators; e.g., age, response rates), where the random effects model was calculated for each level of the moderator. The observed differences in estimates

were assessed for statistical significance using a chi-squared distribution as it represents a widely used and accepted statistical method, aligns with the assumptions of the random-effects model, and is compatibility with analysing moderator effects, statistical inference, hypothesis testing principles (Sun et al., 2018).

All meta-analytic syntheses report confidence intervals and forest plots display the associated prediction interval. The 95% confidence interval for each synthesis represents the range of scores containing the true effect (internal reliability). This confidence interval accounts for variation due to sample size differences and random error. On the other hand, the prediction interval represents the 95% confidence interval for the published literature, considering between-study variations such as population characteristics, study design, test-taking environment, and more. As a result, the prediction interval will always be at least as large, and often larger, than the 95% confidence interval for the synthesis.

Additional Analyses

Methodological quality effects were undertaken by comparing those rated '*Low'* and '*Any'* (Unclear and High) risk of bias. Sensitivity analyses involved 'leave-one-out' assessments by systematically deleting one-to-one reliability coefficients to identify the disproportionately influential studies on the collective effect sizes. Publication bias was assessed using funnel plots showing the magnitude of a study's alpha estimates (i.e., influence) and deviation from the meta-analytic average (i.e., discrepancy). The trim and fill method was used to correct the effects of publication bias to re-compute the effect sizes until the plots were symmetrical, creating mirrored representations and correcting variance without altering the point estimate (Duval & Tweedie, 2000). In the trim and fill method, the adjusted effect size reduces variability, thereby narrowing the confidence interval. Orwin's (1983) Failsafe Number

calculated how many studies with non-significant results would reduce the overall meta-analytical estimate to below minimally interpretable values (α=.70, Nunnally, 1975).

Software

All analyses were performed in R (R Core Team, 2021) with the RStudio GUI (v.1.4.) (RStudio Team, 2021) and metafor package (Viechtbauer, 2010).

Results

Results of the Study Selection Process

Figure 1 illustrates the systematic search from identification to screening, eligibility, and inclusion. The search yielded 1,295 records and removed 863 duplicates and 121 after screening the title. Of the 311 remaining records screened by abstract and 162 full-text reviews, 10 papers did not report any reliability data despite using the MIES. A final sample of 42 unique sources did report reliability data for the MIES, mainly Cronbach's Alpha. Zerach and Levi-Belz (2022) was chosen as it represented multiple papers reporting on a larger overall sample. Several papers reported on the same study using data collected from the same sample which were combined as the '*Veterans Metrics Initiative (VMI) 2020*' (Chesnut et al., 2020; Maguen et al., 2020a, 2020b; Nillni et al., 2020; Richardson et al., 2020). The reason for combining was that each source provided unique and relevant information about the study's outcomes and methodology which together could fully inform the review's selection and risk of bias criteria.

Figure 1

REGEMA Flow Diagram of Selected Studies (Adapted from Sánchez-Meca et al. (2021))

Study Characteristics

Of the 42 studies reporting Cronbach's Alpha, there were 29 for Full-Scale, 22 Betrayal, 19 Transgression-Self, and 19 Transgression-Other, Sub-Scales. Primary study characteristics are reported in Table 2, and alpha coefficients in Table 3. In all 42 records, there were 34,734 participants with an average age of 38.9 years (SD=9.99), 65.4% Males (SD=29.10%), and 68.9% White/Caucasian ethnicity (SD=17.62%). Most studies were US-based (73.8%), and in Military (61.9%) and Community (73.8%) samples. These proportions were similar across Full-Scale and Sub-Scales. The Full-Scale included 30,423 participants, Transgression-Self and Transgression-Other included 28,287 participants, and Betrayal included 29,572 participants.

Table 2

Characteristics of Primary Studies for Alpha Coefficients

Table 3

Alpha Coefficients of Primary Studies for MIES at Full-Scale and Sub-Scales (Transgression-Self, Transgression-Other, Betrayal)

Risk of Bias

The following section includes the ratings of Risk of Bias. Table 4 reports each of the paper's ratings; their corresponding Study Number (#) is referenced below for readability.

Table 4

Ratings of Risk of Bias. A Black Background Indicates a High Risk of Bias, Grey Marks an Unclear Risk of Bias, and White is a Low Risk of Bias

Selection Bias. Selection bias was high, with 26 rated high, followed by 10 unclear, and six low risk. Those rated unclear provided limited information about response rates (4-6, 15, 28, 31), screened participants (though did not exclude them) using MI or trauma-based questionnaires (12, 21, 30), or were recruited via other studies (14). Those rated high were due to the inclusion of non-US military samples (1-3, 7-8, 10-11, 16, 18-20, 22-25, 32, 34-37, 39, 42) or included selectively screened samples (9, 17, 26, 33).

Performance Bias. Performance bias was low, with 29 rated low, 10 unclear, and three high risk. The studies rated high were due to participants being asked to elaborate on their responses including about event details, associated feelings, and impacts on self-perception, relationships, and behaviours (10, 16, 19). Those rated unclear were due to vague administration details (3, 6, 14, 18, 28, 34), or due to differences in how participants were renumerated (e.g., some paid and others not) (5, 17, 40).

Detection Bias. Detection bias was unclear, with 23 rated unclear, 13 low, and six high risk.

Unclear ratings were due to actual or unclear word changes or language adaptations without validation (2, 10, 16, 18, 20, 23-25, 33-34, 36-37, 39, 42), mixed or repeated assessments (3, 14, 18, 27, 29, 36, 42) provision of screening, vignettes or instructions (7, 19, 30, 35), use of screening scales (30), or partial analyses (28). Studies rated high risk were due to items exclusion or splitting (8, 11, 31-32) or rating scale changes (1, 8, 22).

Statistical Bias. Statistical bias was low, with 21 rated low, 18 unclear, and three high risk. Unclear ratings were assigned due to attrition or incomplete data (5-20%) (1-2, 4-5, 7, 15, 22, 24-25, 28, 31-32, 37 40-41) or model changes to fit the data (29). High risk ratings related to missing data (>20%) (30) or select analysis points despite attrition (18).

Reporting Bias. Unclear ratings were given following limited descriptions of probable word changes due to non-military samples or administration procedures (1, 18-19, 24, 31, 34-35, 39, 42), missing or inconsistent data categories and interpretation or statistics in the text, tables, charts, and associated publications (2, 4-6, 8-9, 11-12, 20, 23, 36, 41), or referencing missing sources (14). Those rated high risk were due to probable word changes plus missing or inconsistent reporting of metrics, participant numbers, reimbursement details (22, 25, 28, 30, 40), or not reporting limitations (32).

Generalisability bias. Generalisability was low, with 40 rated low, and 2 unclear due to sample sizes below n<50 (6, 16). It's worth noting the eligibility criteria included studies with sample sizes above n>10, so a lack of high-risk ratings likely reflects the narrow n=10-30 threshold.

Mean Reliability and Heterogeneity

Figure 2 shows the forest plots summarising study results and heterogeneity. In keeping with the generally acceptable levels of minimal internal consistency, the null effect was set to α=.70 (Nunnally, 1975). The pooled alpha coefficients were '*Excellent'* for Transgression-Self (α=.92; 95% CI: [.91-.93]), '*Good'* for the Full-scale (α=.88; 95% CI: [.87-.89]), and '*Moderate'* for Transgression-Other (α=.83; 95% CI: [.80-.85]) and Betrayal (α=.82; 95% CI: [.79-.84]) (Cicchetti, 1994; Ponterotto & Ruckdeschel, 2007). Heterogeneity levels were high at Full-scale (I^2=96%, tau^2=.0006, p<.01) and Transgression-Self (I^2=96%, tau^2=.0003, p<.01), Transgression-Other (I^2=97%, tau^2=.0035, p<.01), and Betrayal (I^2=96%, tau^2=.0023, p<.01) indicating that alpha estimates were likely biased by uncontrolled or confounding factors, likely due to factors other than chance.

Figure 2 illustrates the high heterogeneity across scales. The adjusted prediction intervals are represented in the forest plots as a red line reporting (95% CI: [.83-.94]) for Full-scale, (95% CI: [.88-.93]) Transgression-Self, (95% CI: [.70-.96]) Transgression-Other, and (95% CI: [.71-.92]) Betrayal. For Full-scale and Transgression-Self, most sources centred around the average estimate reflecting the narrow confidence intervals. Chaplo et al. (2019) (α=.70) differed from the average for Transgression-Self, as did Senger et al. (2022) (α=.63) and Ogle et al. (2018) (α=.66) for Transgression-Other. Haight et al. (2017) reported wider confidence intervals for Betrayal (95% CI: α=[.50-.88]), indicating higher intra-sample variability, perhaps reflecting a smaller sample size (n=32). Although these studies represented outliers, all were retained following sensitivity analyses which revealed negligible effects on the overall estimate (see 'Leave-one-out' Analyses).

Figure 1

Forest Plots of Total Alpha Coefficients at Full-scale and Sub-scale Levels

Moderator Analyses

Subgroup analyses of study-level covariates (Steen, 2023a, Supplementary Table 2) reported significant differences at Full-Scale with non-modified items (α=.90), Military samples (α=.90), and US-based studies (α=.89) showing higher values than modified items (α=.84) (p=.007), non-Military (α=.85) (p<.001), and non-US-based studies (α=.85) (p=.036), resulting in category changes from '*Excellent*' to '*Good*'. For Transgression-Self, Military (α=.92) samples reported higher alpha coefficients than non-Military (α=.88) (p=.021), resulting in category changes from '*Excellent*' to '*Good*'. For Transgression-Other, online (α=.86) and paid (α=.82) reported higher estimates than not online (α=.80) (p=.017) and partial

(α=.89) and not paid (α=.77) (p=.003), resulting in category changes from '*Good*' to '*Moderate'* and '*Moderate/Good'* to '*Fair'*. There were no significant differences for Betrayal.

Meta-regression analyses (Steen, 2023a, Supplementary Table 3) reported positive associations with proportions of Males (β=.048, p=.015, R^2=.00%) and those Married (% Currently) (β=.083, p=.044, R^2=6.52%), and negative associations for Education (% College/University) (β=-.068, p=.002, R^2=58.74%), and Depression (β=-.099, p<.001, R^2=82.84%) at Full-Scale. Transgression-Self reported negative associations with the Response rate (β=-.050, p=.039, R^2=8.82%) and proportions of Education (% College/University) (β=-.072, p=.026, R^2=.00%) and PTSD (β=-.040, p=.038, R^2=41.47%) but positive associations with proportions of Males (β=.035, p=.038, R^2=.00%). Transgression-Other reported positive associations with Time in Service (Years) (β=.022, p<.001, R^2=98.86%) and Combat Exposure (%) (β=.181, p<.001, R^2=81.75%) and negative associations with proportions of Males (β=-.097, p=.023, R^2=22.53%). Betrayal showed negative associations with proportions of Ethnicity (% White/Caucasian) and Depression with decreases of β=-.178 (p=.034, R^2=15.17%) and β=-.871 (p<.001, R^2=62.57%) respectively.

Sensitivity Analyses

'Leave-one-out' Analyses

'Leave-one-out' analyses were undertaken to identify disproportionately influential studies by assessing weighted average effect size (i.e., influence) and heterogeneity (i.e., discrepancy) changes with individual records removed. All adjustments equated to approximately ≤1% changes relative to the uncorrected estimates.

Publication Bias

Funnel plots indicated evidence of publication bias at Full-Scale but not Sub-Scales. For the Full-Scale, seven studies were added to the right-side of the plot using the trim and fill method, increasing the estimate by 1.70% to α=.90 (95% CI: [.88-.91]) (Steen, 2023a, Supplementary Figure 1).

Orwin's Failsafe Number

Orwin's (1983) Failsafe Number indicated that 27 studies with an average alpha coefficient of α=.50 or 531 of α=.69 would be required to reduce the observed value for the Full-Scale from α=.88 to α=.70. For the Sub-Scales, these values were 20 studies of α=.50 or 410 of α=.69 for Transgression-Self, 12 studies of α=.50 or 239 of α=.69 for Transgression-Other, and 13 studies of α=.50 or 256 of α=.69 for Betrayal. This indicates the observed values are relatively protected against publication bias and future influential and discrepant studies.

Risk of Bias Effects

Methodological quality effects comparing those rated '*Low'* and '*Any'* (Unclear and High) risk of bias (Steen, 2023a, Supplementary Table 4) found '*Any*' reported higher estimates at Full-Scale for Selection ($p<.001$) and Detection ($p\leq.001$), adjusting ratings from '*Excellent'* to '*Good'*. For Transgression-Self, '*Any*' reported higher estimates for Performance ($p=.025$) and Reporting ($p=.011$) while '*Low*' was higher for Detection ($p=.008$), but all had no impact on rating categories of '*Excellent*'. For Transgression-Other, '*Low*' was higher for Performance ($p=.023$), adjusting ratings from '*Moderate'* to '*Fair'*. There were no significant differences for Betrayal. All comparisons generally remained above the '*Excellent*' or '*Good*' categories, indicating these factors had little impact clinically, thus supporting the internal consistency across different contexts. However, when methodological factors mentioned earlier are present,

the utilisation and interpretation of Transgression-Other should be considered. Based on quality criteria, using non-military populations, making word changes, employing selective screening, and inconsistent administration, as well as utilising less confidential administration formats and encouraging respondents to elaborate on their responses, it is likely to yield lower estimates. These findings underscore the importance of evaluating reliability estimates across all populations, particularly those with non-validated characteristics.

Comparison of Inducing/Omitting and Estimating Studies

The reliability induction rate was 19.2%, with 80.8% reporting estimations based on the current sample, and 9.6% by induction and 9.6% by omission. No statistically significant differences were found for any of the MIES test score comparisons (M; SD) ($p>.211$), Age (M; SD) ($p>.069$) or Ethnicity (% White/Caucasian) ($p=.741$), supporting the meta-analytic estimate's reliability. However, there were differences in Gender (% Male) ($t(35.9)=3.42$, $p=.002$) with studies Inducing/Omitting (M: 85.1%; SD: 10.8%) reporting higher proportions than those Estimating (M: 65.4%; SD: 29.1%), indicating a bias in reporting the internal consistencies between studies.

Data Set

See Steen (2023b) Supplementary Materials.

Discussion

Summary of Results

The meta-analysis supports MIES as an internally consistent tool for assessing PMIEs and associated distress across settings and populations, despite observed heterogeneity. Some indications of publication bias were found but were limited, and estimates remained stable

against future publications. All estimates exceeded the recommended alpha value (α=.70) (Nunnally, 1975), including Full-Scale (α=.88) and Sub-Scales (α=.82-.92). The review is limited by the absence of other reliability and validity metrics, preventing a definitive conclusion about the MIES as a psychometrically sound tool beyond its original design studies (Bryan et al., 2016; Nash et al., 2013). Bias within studies was mostly rated as low (48.8%), followed by unclear (33.7%), and high (17.5%). Despite varying bias ratings, all studies were included due to the limited number available, providing an illustrative summary of the literature. The alpha estimates generally remained at acceptable levels ('*Excellent*' to '*Good*'), despite statistically significant differences between bias categories. The sample mainly consisted of homogeneous English-speaking, male US military personnel, indicating low diversity. Moderator analyses revealed significant findings, particularly in comparisons between military and non-military groups. Mental health and socio-demographic factors (e.g., Depression, PTSD, Gender, Ethnicity, Education) had variable effects across Full-Scale and Sub-Scales. Finally, the reliability induction rate indicated differences between samples regarding the proportion of Males but not MIES test scores, age, or proportion of White/Caucasian ethnicities, supporting the meta-analytic estimation's reliability.

Limitations

Many studies did not prioritise reporting the psychometric properties of MIES, limiting the available data. The findings primarily relied on Cronbach's Alpha as an indicator of internal consistency, which may skew the overall picture since studies with lower alpha values (α<.70) may have not been published. Additionally, 10 papers that used the MIES did not provide any reliability data, further impacting generalisability, although tests of reliability induction suggested limited effects on test score estimations. The high heterogeneity observed in the studies hampers the interpretability and generalisability of the findings. The moderator

analyses of subgroup comparisons may lack statistical power due to small sample sizes, affecting accurate interpretations. Some publications may have been missed if they did not mention the MIES in titles or abstracts, but a full-text review was conducted where this was unclear. Non-English manuscripts (k=17) were excluded, potentially excluding translations of the MIES and favouring studies with higher English fluency. Although there were no significant differences between clinical and community samples, these populations may naturally have higher figures compared to non-treatment-seeking general populations. The decisions on study inclusion and bias ratings relied on the subjective interpretations of the corresponding author (SS), with joint and inter-rater reviews conducted for consensus checking in some cases. Certain forms of response bias, such as exaggerated or malingered presentations, could not be determined and may have affected the data.

Implications for Practice

Psychometric properties are important for selecting suitable assessment tools in clinical and research settings. The estimates of the MIES (α=.82-.92) compare favorably with other MI assessment tools across Full (α=.82-.95) and Sub-Scales (α=.56-.98) (Chaplo et al., 2019; Currier et al., 2018; Fani et al., 2021; Hoffman & Nickerson, 2021; Koenig et al., 2018). These estimates are supported by diverse administration contexts (timings, items, assessors, settings), affirming the tool's applicability across settings and populations. The review expands the assessment of psychometric properties beyond its initial design studies (Bryan et al., 2016; Nash et al., 2013), demonstrating the MIES' internal consistency in broader settings. However, this review does not address the tool's structural validity or its accuracy in measuring MI across contexts. The utility of the MIES in measuring the MI concept, which is a topic of ongoing debate, is not addressed in these findings. Future research should focus on operationalising the concept and aligning the tool with evolving definitions.

The large heterogeneity suggests that the MIES is influenced by study and sample characteristics, limiting its generalisability. Significant differences were observed in non-US and non-military settings with item modifications, which are reflected in the significant differences for Selection and Detection biases. The development and understanding of MI within US military settings, including the design of the MIES for this context (Nash et al., 2013), may explain these differences. Researchers and professionals working with military personnel have developed interventions for MI, such as Acceptance and Commitment Therapy for MI (ACT-MI) (Borges, 2019; Farnsworth et al., 2017). Clearer criteria and an improved understanding of MI may lead to more accurate assessments and reduced variations in item interpretations. In non-military settings, where MI is a relatively new concept, item changes vary (e.g., '*animal shelter*' '*colleagues*' or '*healthcare or public health organizations*') (Andrukonis & Protopopova, 2020; Haight et al., 2017; Khan et al., 2021), naturally creating inconsistencies for interpretation. These findings emphasise the need for researchers and clinicians to be cautious when using the MIES in non-US military settings and consider using tools designed and validated for specific contexts. However, the MIES can still be a viable alternative as rating categories generally remained above *'Good*'. It is recommended that researchers and clinicians integrate psychoeducation to enhance the interpretation of the MIES items, as improved understanding and education about MI could contribute to improved assessments.

Possible sources of heterogeneity included gender, ethnicity, marital status, education level, depression, PTSD, assessment format, and payment. These variables may have influenced the interpretation of items related to the relatively recent and abstract concept of MI and the MIES' generalised, multidimensional, and non-temporal features. Research has found gender differences in MI following PMIEs, indicating that gender plays a role in interpretation

(Maguen et al., 2020a; Maguen et al., 2022; Nieuwsma et al., 2022; Webb et al., 2023). Ethnicity influenced the Betrayal Sub-Scale, possibly reflecting broader social empowerment factors (Nieuwsma et al., 2022). The influence of marital status on MI-Symptoms is mixed potentially explaining the variability observed (Forkus et al., 2021; Khan et al., 2021; Zearch & Levi-Belz, 2022). Education levels might reflect different capacities to interpret MI's abstract concept which could reflect the lack of consensus in definitions (Koenig et al., 2019) and the need for consistent assessment procedures and focus on unidimensional MI factors (Williamson et a., 2021). Comorbid mental health conditions, such as PTSD and depression, are associated with MI but vary in their impact and therefore likely on item interpretation (Maguen et al., 2020a; Williamson et al., 2022). It is recommended these factors be robustly assessed and reported when using the MIES to identify differences in interpretation.

The Transgression-Other Sub-Scale reported higher estimates for online assessment formats involving payment and an increased time in service and combat exposure. As this relates to observing transgressive acts committed by others, it's plausible that increased exposure increases its likelihood and so respondents answer more consistently to abstract items with concrete experiences. The assessment format may have encouraged respondents to be more open about these transgressive acts (Williamson et a., 2021). Online applications can increase accessibility and may be more appropriate for assessing others' transgressions, particularly in the context of military settings with increased combat exposure. Based on these findings, online assessment formats involving payment might be advisable in military contexts to enhance internal consistency.

Inconsistencies in the interpretation of others' transgressive acts within non-military samples likely stem from the range of transgressive behaviours observed, which vary according to the

unique contextual factors characteristic of different population groups. It is important to acknowledge that the interpretation of transgressive acts committed by others is not equivalent across all non-military samples. For instance, the manner in which transgressive acts are perceived and understood by animal shelter employees (Andrukonis & Protopopova, 2020), which can be contingent upon their involvement in euthanising animals, differs from the perspective of the general population scoring others' transgression higher when they have a personal connection to someone afflicted by COVID-19 during the pandemic (Khan et al., 2021). Consequently, it becomes imperative to consider the population-specific variations in the conceptualisation and interpretation of transgressive acts committed by others. It is also necessary that researchers and practitioners exercise caution when inferring internal consistencies across contextual settings. In the development of assessment tools and the formulation of intervention strategies, there is a pressing need to re-contextualise the item ratings in alignment with the unique characteristics and perspectives of the specific populations. Recognising and accounting for these population-specific distinctions can ensure that methods and interventions remain contextually relevant and effectively tailored to the diverse interpretations and experiences within distinct groups.

Only Transgression-Other had a '*Fair*' rating in moderator analyses, possibly due to its low item count (k=2). Cronbach's Alpha tends to increase with more items, but assessment tools must balance internal consistency with the risk of redundancy (Hair, 2014). Note that alpha estimates above α=.90, like Transgression-Self, may contain duplicate items, despite being categorised as '*Excellent*' (Cicchetti, 1994; Ponterotto & Ruckdeschel, 2007). The MIES initially had a two-factor model but was later revised by Bryan et al. (2016) into Self and Other categories, which was commonly used in this review. The low item count and general wording of Transgression-Other may lead to different interpretations, particularly regarding the

relationship between PMIEs and MI-Symptoms when observing others' acts. Self-transgressions may be more readily identifiable, leading to consistent interpretations. Respondents may vary in disclosing others' transgressions, as seen in the assessment format effects. Refining the Transgression-Other Sub-Scale may be necessary for better consistency across contexts, and confirmatory factor analyses would be valuable when using and interpreting the MIES Sub-Scales.

Implications for Future Research

The analyses suggest that methodological quality ratings do not substantially change alpha estimate categories apart from Transgression-Other. The MIES demonstrates good internal consistency across studies, but its psychometric properties are limited. While pooled alpha estimates support the use of the MIES, other factors such as reliability, validity, and specific populations and outcomes should also be considered when choosing assessment tools. Reporting alternative reliability and validity metrics would enhance the field of MI assessment. Consistency in concepts and methods would also enable benchmarking for comparisons. The review could not develop overall average MIES scores due to inconsistencies in how studies reported their scores. The field of MI assessment is diverse, not only within a single tool but more broadly. New measurement approaches offer opportunities but lack validity and reliability testing (Hodgson et al., 2021; Williamson et al., 2020; Yeterian et al., 2019).

Summary Conclusions

This review examined the pooled alpha estimates of the MIES across studies, revealing evidence of its internal consistency in diverse populations and contexts. Despite high heterogeneity and tool adjustments, particularly for non-US military samples, the MIES maintained acceptable levels from a clinical perspective. The review expands the analysis of

the MIES psychometric properties beyond its original design (Bryan et al., 2016; Nash et al., 2013) and highlights its consistency across various applications of measurement, including settings, populations, and administration. It provides insights for professionals considering the MIES as an assessment tool and emphasises the importance of reporting all relevant psychometric properties to enhance its reliability and validity across different settings. Additional assessment procedures are recommended including assessing and reporting on gender, ethnicity, marital status, education level, depression, PTSD, assessment format, and payment to identify possible differences in interpretation. Integrating psychoeducation might help improve the understanding and interpretation of MIES items, thereby enhancing assessments.

Funding

This review was completed as part of a Clinical Psychology Doctorate and no funding was received from any organisations.

Acknowledgements

This manuscript is adapted from a PhD thesis that is not under embargo.

References

*References marked with an asterisk indicate studies included in the meta-analysis.

*Amsalem, D., Lazarov, A., Markowitz, J. C., Naiman, A., Smith, T. E., Dixon, L. B., & Neria, Y. (2021). Psychiatric symptoms and moral injury among US healthcare workers in the COVID-19 era. *BMC Psychiatry*, *21*(1), 546.

*Andrukonis, A., & Protopopova, A. (2020). Occupational Health of Animal Shelter Employees by Live Release Rate, Shelter Type, and Euthanasia-Related Decision. *Anthrozoös*, *33*(1), 119–131.

Bender, R., Friede, T., Koch, A., Kuß, O., Schlattmann, P., Schwarzer, G., … & Skipka, G. (2018). Methods for evidence synthesis in the case of very few studies. Research Synthesis Methods, 9(3), 382-392. https://doi.org/10.1002/jrsm.1297

*Bhalla, A., Allen, E., Renshaw, K., Kenny, J., & Litz, B. (2018). Emotional numbing symptoms partially mediate the association between exposure to potentially morally injurious experiences and sexual anxiety for male service members. *Journal of Trauma & Dissociation : The Official Journal of the International Society for the Study of Dissociation (ISSD)*, *19*(4), 417–430.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd Ed.). John Wiley & Sons.

*Bryan, C. J., Bryan, A. O., Anestis, M. D., Anestis, J. C., Green, B. A., Etienne, N., Morrow, C. E., & Ray-Sannerud, B. (2016). Measuring moral injury: Psychometric properties of the moral injury events scale in two military samples. *Assessment*, *23*(5), 557–570.

*Cameron, A. Y., Eaton, E., Brake, C. A., & Capone, C. (2020). Moral injury as a unique predictor of suicidal ideation in a veteran sample with a substance use disorder. *Psychological Trauma : Theory, Research, Practice and Policy*, 13(8), 856-860.

*Chaplo, S. D., Kerig, P. K., & Wainryb, C. (2019). Development and validation of the moral injury scales for youth. *Journal of Traumatic Stress*, *32*(3), 448–458.

Chen, H., Manning, A., & Dupuis, J. (2012). A method of moments estimator for random effect multivariate meta-analysis. Biometrics, 68(4), 1278-1284. https://doi.org/10.1111/j.1541-0420.2012.01761.x

*Chesnut, R. P., Richardson, C. B., Morgan, N. R., Bleser, J. A., Perkins, D. F., Vogt, D., Copeland, L. A., & Finley, E. (2020). Moral Injury and Social Well-Being: A Growth Curve Analysis. *Journal of Traumatic Stress*, *33*(4), 587–597.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290.

Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, *64*(3), 391–418.

Currier, J. M., Farnsworth, J. K., Drescher, K. D., McDermott, R. C., Sims, B. M., & Albright, D. L. (2018). Development and evaluation of the E xpressions of M oral I njury S cale—M ilitary V ersion. *Clinical Psychology & Psychotherapy*, *25*(3), 474–488.

*Dale, L. P., Cuffe, S. P., Sambuco, N., Guastello, A. D., Leon, K. G., Nunez, L. V., Bhullar, A., Allen, B. R., & Mathews, C. A. (2021). Morally Distressing Experiences, Moral Injury, and Burnout in Florida Healthcare Providers during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, *18*(23), Article 23.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

DerSimonian, R. and Laird, N. (2015). Meta-analysis in clinical trials revisited. Contemporary Clinical Trials, 45, 139-145. https://doi.org/10.1016/j.cct.2015.09.002

Duval S & Tweedie R (2000), Trim and Fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455463.

*Evans, W. R., Szabo, Y. Z., Stanley, M. A., Barrera, T. L., Exline, J. J., Pargament, K. I., & Teng, E. J. (2018). Life satisfaction among veterans: Unique associations with morally injurious events and posttraumatic growth. *Traumatology*, *24*(4), 263–270.

*Fani, N., Currier, J. M., Turner, M. D., Guelfo, A., Kloess, M., Jain, J., Mekawi, Y., Kuzyk, E., Hinrichs, R., Bradley, B., Powers, A., Stevens, J. S., Michopoulos, V., & Turner, J. A. (2021). Moral injury in civilians: Associations with trauma exposure, PTSD, and suicide behavior. *European Journal of Psychotraumatology*, *12*(1), 1965464.

*Feinstein, A., Pavisian, B., & Storm, H. (2018). Journalists covering the refugee and migration crisis are affected by moral injury not PTSD. *JRSM Open*, *9*(3), 2054270418759010.

*Forkus, S. R., Breines, J. G., & Weiss, N. H. (2019). Morally injurious experiences and mental health: The moderating role of self-compassion. *Psychological Trauma : Theory, Research, Practice and Policy*, *11*(6), 630–638.

*Forkus, S. R., Schick, M. R., Goncharenko, S., Thomas, E. D., Contractor, A. A., & Weiss, N. H. (2021). The moderating role of emotion dysregulation in the relation between potentially morally injurious experiences and alcohol misuse among military Veterans. *Military Psychology*, *33*(1), 41–49.

*Frankfurt, S. B., DeBeer, B. B., Morissette, S. B., Kimbrel, N. A., Bash, H. L., & Meyer, E. C. (2018). Mechanisms of Moral Injury Following Military Sexual Trauma and Combat in Post-9/11 U.S. War Veterans. *Frontiers in Psychiatry*, *9*, 520.

*Griffin, B. J., Williams, C. L., Shaler, L., Dees, R. F., Cowden, R. G., Bryan, C. J., Litz, B., Purcell, N., Burkman, K., & Maguen, S. (2020). Profiles of moral distress and associated outcomes among student veterans. *Psychological Trauma: Theory, Research, Practice and Policy*, *12*(7), 669–677.

Guolo, A. and Varin, C. (2015). Random-effects meta-analysis: the number of studies matters. Statistical Methods in Medical Research, 26(3), 1500-1518. https://doi.org/10.1177/0962280215583568

*Haight, W., Sugrue, E., Calhoun, M., & Black, J. (2017). "Basically, I look at it like combat": Reflections on moral injury by parents involved with child protection services. *Children and Youth Services Review*, *82*, 477–489.

Hair, J. F. (Ed.). (2014). *Multivariate data analysis* (7. ed.,). Pearson.

*Held, P., Klassen, B. J., Steigerwald, V. L., Smith, D. L., Bravo, K., Rozek, D. C., Horn, R. V., & Zalta, A. (2021). Do morally injurious experiences and index events negatively impact intensive PTSD treatment outcomes among combat veterans? *European Journal of Psychotraumatology*, *12*(1), 1877026.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560.

*Hines, S. E., Chin, K. H., Glick, D. R., & Wickwire, E. M. (2021). Trends in Moral Injury, Distress, and Resilience Factors among Healthcare Workers at the Beginning of the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, *18*(2).

Hodgson, T. J., Carey, L. B., & Koenig, H. G. (2021). Moral Injury, Australian Veterans and the Role of Chaplains: An Exploratory Qualitative Study. *Journal of Religion and Health*, *60*(5), 3061–3089.

Hoffman, J., & Nickerson, A. (2021). The Impact of Moral-Injury Cognitions on Psychological Outcomes in Refugees: An Experimental Investigation. *Clinical Psychological Science*, 21677026211039516.

*Houle, S. A., Vincent, C., Jetly, R., & Ashbaugh, A. R. (2021). Patterns of distress associated with exposure to potentially morally injurious events among Canadian Armed Forces service members and Veterans: A multi-method analysis. *Journal of Clinical Psychology*, *77*(11), 2668–2693.

Jinkerson, J. D. (2016). Defining and assessing moral injury: A syndrome perspective. *Traumatology*, *22*(2), 122.

*Khan, A. J., Nishimi, K., Tripp, P., Maven, D., Jiha, A., Woodward, E., Inslicht, S., Richards, A., Neylan, T. C., Maguen, S., & O'Donovan, A. (2021). COVID-19 related moral injury: Associations with pandemic-related perceived threat and risky and protective behaviors. *Journal of Psychiatric Research*, *142*, 80–88.

*Kinney, A. R., Gerber, H. R., Hostetter, T. A., Brenner, L. A., Forster, J. E., & Stephenson, R. O. (2022). Morally injurious combat events as an indirect risk factor for postconcussive symptoms among veterans: The mediating role of posttraumatic stress. *Psychological Trauma: Theory, Research, Practice, and Policy*, 15(1), 144-152.

Koenig, H. G., Ames, D., Youssef, N. A., Oliver, J. P., Volk, F., Teng, E. J., Haynes, K., Erickson, Z. D., Arnold, I., & O'Garo, K. (2018). The moral injury symptom scale-military version. *Journal of Religion and Health*, *57*(1), 249–265.

Koenig, H. G., Youssef, N. A., & Pearce, M. (2019). Assessment of moral injury in veterans and active duty military personnel with PTSD: A review. *Frontiers in Psychiatry*, *10*, 443.

Langan, D., Higgins, J., Jackson, D., Bowden, J., Veroniki, A., Kontopantelis, E., … & Simmonds, M. (2018). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods, 10(1), 83-98. https://doi.org/10.1002/jrsm.1316

*Lee, H., Aldwin, C. M., & Kang, S. (2020). Do different types of war stressors have independent relations with mental health? Findings from the Korean Vietnam Veterans Study. *Psychological Trauma: Theory, Research, Practice, and Policy*, 14(5), 840-848.

*Levi-Belz, Y., & Zerach, G. (2022). The wounded helper: Moral injury contributes to depression and anxiety among Israeli health and social care workers during the COVID-19 pandemic. *Anxiety, Stress, & Coping*, *0*(0), 1–15.

*Litam, S. D. A., & Balkin, R. S. (2020). Moral injury in health-care workers during COVID-19 pandemic. *Traumatology*. Advance online publication.

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.

*Maftei, A., & Holman, A.-C. (2021). The prevalence of exposure to potentially morally injurious events among physicians during the COVID-19 pandemic. *European Journal of Psychotraumatology*, *12*(1), 1898791.

*Maguen, S., Griffin, B. J., Copeland, L. A., Perkins, D. F., Finley, E. P., & Vogt, D. (2020a). Gender differences in prevalence and outcomes of exposure to potentially morally injurious events among post-9/11 veterans. *Journal of Psychiatric Research*, *130*, 97–103.

*Maguen, S., Griffin, B. J., Copeland, L. A., Perkins, D. F., Richardson, C. B., Finley, E. P., & Vogt, D. (2020b). Trajectories of functioning in a population-based sample of veterans: Contributions of moral injury, PTSD, and depression. *Psychological Medicine*, 1–10.

*Maguen, S., Griffin, B. J., Vogt, D., Hoffmire, C. A., Blosnich, J. R., Bernhard, P. A., Akhtar, F. Z., Cypel, Y. S., & Schneiderman, A. I. (2022). Moral injury and peri- and post-military suicide attempts among post-9/11 veterans. *Psychological Medicine*, 1–10.

*Maguen, S., Nichter, B., Norman, S. B., & Pietrzak, R. H. (2021). Moral injury and substance use disorders among US combat veterans: Results from the 2019-2020 National Health and Resilience in Veterans Study. *Psychological Medicine*, 1–7.

*Martin, R. L., Houtsma, C., Bryan, A. O., Bryan, C. J., Green, B. A., & Anestis, M. D. (2017). The impact of aggression on the relationship between betrayal and belongingness among U.S. military personnel. *Military Psychology*, *29*(4), 271–282.

McEwen, C., Alisic, E., & Jobson, L. (2020). Moral injury and mental health: A systematic review and meta-analysis. *Traumatology*, 27(3), 303–315.

*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.

*Nieuwsma, J. A., Brancu, M., Wortmann, J., Smigelsky, M. A., King, H. A., & Meador, K. G. (2021). Screening for moral injury and comparatively evaluating moral injury measures in relation to mental illness symptomatology and diagnosis. *Clinical Psychology & Psychotherapy*, *28*(1), 239–250.

Nieuwsma, J. A., O'Brien, E. C., Xu, H., Smigelsky, M. A., VISN 6 MIRECC Workgroup, HERO Research Program, & Meador, K. G. (2022). Patterns of potential moral injury in post-9/11 combat veterans and COVID-19 healthcare workers. *Journal of General Internal Medicine*, 37(8), 2033-2040.

*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 Veterans' Perinatal Outcomes Following Separation From Military Service. *Journal of Traumatic Stress*, *33*(3), 248–256.

Nunnally, J. C. (1975). Psychometric Theory—25 Years Ago and Now. *Educational Researcher*, *4*(10), 7–21.

Nunnally, J. C., & Bernstein, I. (1993). *Psychometric Theory* (3rd edition). McGraw Hill.

*Ogle, A. D., Reichwald, R., & Rutland, J. B. (2018). Psychological impact of remote combat/graphic media exposure among US Air Force intelligence personnel. *Military Psychology*, *30*(6), 1–11.

Orwin, R. G. (1983). A Fail-Safe N for Effect Size in Meta-Analysis. *Journal of Educational Statistics*, *8*(2), 157–159.

*Papazoglou, K., Blumberg, D. M., Chiongbian, V. B., Tuttle, B. M., Kamkar, K., Chopko, B., Milliard, B., Aukhojee, P., & Koskelainen, M. (2020). The Role of Moral Injury in PTSD Among Law Enforcement Officers: A Brief Report. *Frontiers in Psychology*, *11*, 310.

*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*, 10731911211044198.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An Overview of Coefficient Alpha and a Reliability Matrix for Estimating Adequacy of Internal Consistency Coefficients with Psychological Research Measures. *Perceptual and Motor Skills*, *105*(3), 997–1014.

*Protopopescu, A., Boyd, J. E., O'Connor, C., Rhind, S. G., Jetly, R., Lanius, R. A., & McKinnon, M. C. (2021). Examining the associations among moral injury, difficulties with emotion regulation, and symptoms of PTSD, depression, anxiety, and stress among Canadian military members and Veterans: A preliminary study. *Journal of Military, Veteran and Family Health*, *7*(2), 71–80.

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* https://www.R-project.org

RStudio Team (2021). *RStudio: Integrated Development for R. RStudio (v.1.4.)*, Inc., Boston, MA. http://www.rstudio.com/.

*Richardson, C. B., Chesnut, R. P., Morgan, N. R., Bleser, J. A., Perkins, D. F., Vogt, D., Copeland, L. A., & Finley, E. (2020). Examining the factor structure of the moral injury events scale in a veteran sample. *Military Medicine*, *185*(1–2), e75–e83.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the reporting quality of reliability

generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, 12(4), 516–536.

*Schwartz, G., Halperin, E., & Levi-Belz, Y. (2021). Moral Injury and Suicide Ideation Among Combat Veterans: The Role of Trauma-Related Shame and Collective Hatred. *Journal of Interpersonal Violence*, 088626052110079.

*Senger, A. R., Torres, D., & Ratcliff, C. G. (2022). Potentially morally injurious events as a mediator of the association of gratitude and mindfulness with distress. *Psychological Trauma: Theory, Research, Practice, and Policy*. Online ahead of print.

Shay, J. (1995). *Achilles in Vietnam: Combat Trauma and the Undoing of Character* (Touchstone ed). Pocket Books.

Sidik, K. and Jonkman, J. (2006). A comparison of heterogeneity variance estimators in combining results of studies. Statistics in Medicine, 26(9), 1964-1981. https://doi.org/10.1002/sim.2688

Steen, S. (2023a). *The Internal Consistency of the Moral Injury Event Scale: A Reliability Generalisation Meta-Analysis and Systematic Review, 2021-2022.* Colchester, Essex: UK Data Service. https://dx.doi.org/10.5255/UKDA-SN-856807

Steen, S. (2023b). *A Meta-Analysis of the Internal Consistency of the Moral Injury Event Scale, 2021-2022.* Colchester, Essex: UK Data Service. https://dx.doi.org/10.5255/UKDA-SN-856549

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

*Sugrue, E. P. (2020). Moral Injury Among Professionals in K–12 Education. *American Educational Research Journal*, *57*(1), 43–68.

Sun, J., Freeman, B., & Natanson, C. (2018). Meta-analysis of clinical trials., 317-327. https://doi.org/10.1016/b978-0-12-849905-4.00022-8

*Thomas, E. D., Weiss, N. H., Forkus, S. R., & Contractor, A. A. (2021). Examining the Interaction Between Potentially Morally Injurious Events and Religiosity in Relation to Alcohol Misuse Among Military Veterans. *Journal of Traumatic Stress*, 35(1), 314-320.

*Ulusoy, S., & Celik, Z. (2022). The Silent Cry of Healthcare Workers: A Cross-Sectional Study on Levels and Determinants of Burnout among Healthcare Workers after First Year of the Pandemic in Turkey. *Psychiatry and Clinical Psychopharmacology*, *32*(1), 63–71.

Venazzi, A., Swardfager, W., Lam, B., Siqueira, J. de O., Herrmann, N., & Cogo-Moreira, H. (2018). Validity of the QUADAS-2 in Assessing Risk of Bias in Alzheimer's Disease Diagnostic Accuracy Studies. *Frontiers in Psychiatry*, *0*.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*, 1–48.

Wang, Z., Alzuabi, M., Mustafa, R., Falck-Ytter, Y., Dahm, P., Sultan, S., … & Murad, M. (2023). Different meta-analysis methods can change judgements about imprecision of effect estimates: a meta-epidemiological study. BMJ Evidence-Based Medicine, 28(2), 126-132. https://doi.org/10.1136/bmjebm-2022-112053

Webb, E. L., Morris, D. J., Sadler, E., MacMillan, S., Trowell, S., & Legister, A. (2023). Predictors of Moral Injury in Secure Mental Healthcare Workers: Examining a Role for Violence and Restrictive Practices Through an Intersectional Lens. *Journal of Forensic Psychology Research and Practice*, 0(0), 1–19.

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., Bossuyt, P. M. M., & QUADAS-2 Group. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–536.

Williamson, V., Murphy, D., Stevelink, S. A. M., Jones, E., & Greenberg, N. (2020). *Experiences of moral injury in UK military veterans*. London: King's College London.

Williamson, V., Murphy, D., Castro, C., Vermetten, E., Jetly, R., & Greenberg, N. (2020). Moral injury and the need to carry out ethically responsible research. *Research Ethics*, 174701612096974.

*Wisco, B. E., Marx, B. P., May, C. L., Martini, B., Krystal, J. H., Southwick, S. M., & Pietrzak, R. H. (2017). Moral injury in U.S. combat veterans: Results from the national health and resilience in veterans study. *Depression and Anxiety*, 34(4), 340-347.

Yeterian, J. D., Berke, D. S., Carney, J. R., McIntyre-Smith, A., St. Cyr, K., King, L., Kline, N. K., Phelps, A., Litz, B. T., & Consortium, M. I. O. P. (2019). Defining and measuring moral injury: Rationale, design, and preliminary findings from the moral injury outcome scale consortium. *Journal of Traumatic Stress*, *32*(3), 363–372.

*Zerach, G., & Levi-Belz, Y. (2022). Exposure to combat incidents within military and civilian populations as possible correlates of potentially morally injurious events and moral injury outcomes among Israeli combat veterans. *Clinical Psychology & Psychotherapy*, *29*(1), 274–288.

Open Science

We report how we determined our sample size, all data exclusions, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data: The information needed to reproduce all of the reported results are not openly accessible.

Open Materials: I confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology. The data that support the findings of this review are openly available via the UK Data Service: Reference number(s): https://dx.doi.org/10.5255/UKDA-SN-856807; https://dx.doi.org/10.5255/UKDA-SN-856549.

Preregistration of Studies and Analysis Plans: This study was preregistered with an analysis plan on Prospero: Reference number: CRD42021256446.

Open Analytic Code: The scripts, code, and outputs needed to reproduce all of the reported results are not openly accessible.