Supplementary material Table S1. Systematic literature search strategy, exemplified by the search in the MEDLINE and PsycINFO databases

PROM	Searches	Results
Oldenburg Burnout Inventory (OLBI)		
1	(Oldenburg Burnout Inventory or OLBI).mp.[mp=ti, ab, tx, ct, ot nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	303
2	Validation or validity or reliability or psychometric* or equivalence or invariance).mp. [mp=ti, ab, tx, ct, ot, nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	799727
3	1 and 2	77
4	Remove duplicates from 3	68
Copenhagen Burnout Inventory (CBI)		
1	(Copenhagen Burnout Inventory or CBI).mp.[mp=ti, ab, tx, ct, ot nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	1776
2	Validation or validity or reliability or psychometric* or equivalence or invariance).mp. [mp=ti, ab, tx, ct, ot, nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	799727
3	1 and 2	251
4	Remove duplicates from 3	212
Burnout measure (BM)		
1	(Pines Burnout Measure or Pines or BM).mp.[mp=ti, ab, tx, ct, ot nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	26340
2	Validation or validity or reliability or psychometric* or equivalence or invariance).mp. [mp=ti, ab, tx, ct, ot, nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	799727
3	1 and 2	796
4	Remove duplicates from 3	775
Psychologists Burnout Inventory (PBI)		
1	(Psychologists Burnout Inventory or PBI).mp.[mp=ti, ab, tx, ct, ot nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	2347
2	Validation or validity or reliability or psychometric* or equivalence or invariance).mp. [mp=ti, ab, tx, ct, ot, nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	799727
3	1 and 2	306
4	Remove duplicates from 3	265
Maslach Burnout Inventory (MBI)		

1	(MBI or burnout measure or MBS or BM or Maslach Burnout Inventory or MBI dimensions or subscale of the Maslach burnout inventory or Maslach burnout inventory or general Survey or MBI-GS or MBI-HSS or Maslach).mp.[mp=ti, ab, tx, ct, ot nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	37066
2	Validation or validity or reliability or psychometric* or equivalence or invariance).mp. [mp=ti, ab, tx, ct, ot, nm, hw, fx, kf, px, rx, ui, sy, tc, id, tm]	799727
3	1 and 2	2228
4	Remove duplicates from 3	2082

Supplementary table S2. Detailed results of agreement between the authors and reviewers

PROM	Quantitative data measured	Analysis / techniques	Indices and the reviewers' interpretation	Author, year	Results	Results of the comparison between authors’ and reviewers’ interpretation
Maslach Burnout Inventory (MBI)	Reliability (Homogeneity)	Alpha Cronbach, α	α > 0.9, excellent α= 0.8-0.9 & the number of itmes ≤10, good α= 0.8-0.9 & the number of itmes 11-30, just acceptable α= 0.7-0.8 & the number of itmes ≤10, acceptable α=0.6-0.7, questionable α=0.5-0.6, poor α=<0.5, unacceptable	Boles, 2000	For the sample of educators and for frequency EE : 0.89, PA : 0.76 and DP : 0.80	Partial agreement
				Boles, 2000	For the sample of business owners and for frequency EE : 0.90, PA : 0.78 and DP : 0.70	Partial agreement
				Chao, 2011	For frequency EE : 0.91, PA : 0.62 and DP : 0.76	Total agreement
				Gold, 1984	For frequency: EE : 0.88, PA : 0.74, and DP : 0.72 and for intensity: EE : 0.87, DP : 0.79 and PA : 0.75	Total agreement
				Iwanicki, 1981	For frequency EE : 0.90, PA : 0.76 and DP : 0.76 and for intensity EE : 0.89, DP : 0.79 and PA : 0.75	Total agreement
				Kalliath, 2000	For a sample of nurses and not specified whether for frequency or intensity EE : 0.90, PA : NA and DP : 0.76	Disagreement
				Kalliath, 2000	For a sample of laboratory technicians and not specified whether for frequency or intensity: EE : 0.84, PA : XX and DP : 0.75	Disagreement

			Kalliath, 2000	For a sample of managers and not specified whether for frequency or intensity: EE : 0.84, PA : XX and DP : 0.71	Disagreement
			Kim, 2008	For frequency: EE : 0.92, PA : 0.80 and DP : 0.77	Total agreement
			Lahoz, 1989	For frequency EE : 0.90, PA : 0.79 and DP : 0.74 and for intensity: EE : 0.89, DP : 0.79 and PA : 0.75	Total agreement
			Meier, 1984	Not specified whether for frequency or intensity EE : 0.92, PA : 0.80 and DP : 0.77	Disagreement
			Poghosyan, 2009	Cronbach alphas for all countries exceed the critical value of 0.70, except for the depersonalization dimension in Armenia.	Total agreement
			Yadama, 1995	EE: 0.88, DP: 0.80, and PA:0.74	Total agreement
Construct Validity (Factorial Analyses)	Exploratory factorial analyses (EFA) 1) Extraction 2) Rotation	Values ≥ 0.90 was considered to indicate acceptable model fit Values ≥ 0.95 is presently accepted as an indicator of good fit	Brookings, 1985	For frequency 1) Scree test (4) Communality : 0.85 2) Quartimin method % Variance (h2) : 0.85 (EE) 0.92 (PA) 0.66 (DP)	Total agreement
			Gold, 1992	For frequency 1) NA 2) Oblimin : NR	Total agreement
			Iwanicki, 1981	1) Scree test (4) 2) Varimax Method Frequency Intensity Eigenvalues >1 Eigenvalues >1 % Variance : 55 % Variance : 55 2) Oblique Frequency Intensity Eigenvalues >1 Eigenvalues >1 % Variance : 55 % Variance : 55	Total agreement

		Lahoz, 1989	1) NA 2) Varimax Method Frequency Intensity Eigenvalues EE =7.02 Eigenvalues EE = 6.56 Eigenvalues PA = 2.81 Eigenvalues PA = 3.12 Eigenvalues DP = 1.4 Eigenvalues DP = 1.46 % Variance : 51 % Variance : 50.6	Disagreement
		Chao, 2011	EFA investigated an alternative factor structure, a four-factor model dividing the DP dimension into two factors (DP1 –indifference and DP2–rejection) was suggested.	Total agreement
		Holland, 1994	For a sample of teachers, EFA was conducted for two hypothesized dimensions and three hypothesized dimensions model. A close degree of correspondence is noted between both orthogonal and oblique solutions and both within principal components and principal factors approaches.	Disagreement
		Poghosyan, 2009	For a sample of social workers, they began with MBI that is widely used with 22 items and a three-factor structure. They also tested the validity of the revised MBI with 18 items. The new re-specified MBI had a much better fit than the original MBI.	Partial agreement
Confirmatory factorial analyses (CFA)	Values ≥ 0.90 was considered to indicate acceptable model fit Values ≥ 0.95 is presently accepted as an indicator of good fit	Beckstead, 2002	Frequency Communality : 0.449 GFI : 0.78 AGFI : 0.73 CFI : 0.82 RMSEA : 0.09 SRMR : 0.11 X^2 = 452.55, df 206, Null model	Total agreement
		Gold, 1992	GFI : 0.793 AGFI : 0.746 RMSR : 0.177 X^2 = 396.49, df : 206, Modell Null	Disagreement

Kim, 2011	CFI : 0.86 RMSEA : 0.08 Standardized root-mean-square error of approximation X2 = 892, df 206, Model One Factor	Disagreement
Holland, 1994	GFI : 0.777 GFI Model Orthogonal : 0.745 AGFI : 0.726 AGFI Model Orthogonal : 0.696 RMSR : 0.085 RMSR Model Orthogonal : 0.234 X2 = 455.97, df = 206, Null Model X2 model orthogonal = 575.41, df = 212	Disagreement
Chao, 2011	GFI : 0.85 RMSEA : 0.079 AIC : 751.38 X2 = 657.38, df = 206, Null Model	Total agreement
Gold, 1984	Communality: Range H2 Frequency EE : 39-62 (SM: 69) Range H2 Frequency PA : 16-52 Range H2 Frequency DP : 22-66 Range H2 Intensity EE : 38-68 Range H2 Intensity PA : 22-50 Range H2 Intensity DP : 24-63 Varimax rotation Frequency Intensity Eigenvalues EE = 5.8 Eigenvalues EE = 5.41 Eigenvalues PA = 1.08 Eigenvalues PA = 1.23 Eigenvalues DP = 1.93 Eigenvalues DP = 2.55	Disagreement
Brookings, 1985	Using the scree criterion, four components were retained and rotated to oblique simple structure by the quartimin method.	Total agreement
Boles, 2000	They examined the dimensionality of burnout through confirmatory factor analysis (CFA) with LISREL VIII. Only the CFI changed from .88 in the baseline model to .87 in the factor loadings invariant model.	Disagreement

			Lahoz, 1989	Factor analysis (principal factoring) with iteration and an orthogonal (varimax) rotation was used. The three factors accounted for 51.0% of the total variance in the frequency dimension with corresponding eigenvalues of 7.02, 2.81, and 1.40. For the intensity dimension, 50.6% of the total variance was accounted for by the three factors with eigenvalues of 6.56, 3.12, and 1.46.	Total agreement
			Poghosyan, 2009	While the values of the Root Mean Square Error of Approximation (RMSEA) and Bartlett’s Comparative Fit Index (CFI) approach the values that are usually considered acceptable (i.e., RMSE < .06 and CFI > .90, respectively), the RMSEA shows an acceptable fit only in Russia and the CFI value is unacceptable in every country. Moreover, the chi-square statistic indicating the goodness-of-fit in each country suggests an unacceptable fit of model to data in every country.	Disagreement
			Yadama, 1995	The null model has a GFI of 0.79 and an adjusted goodness-of-fit index (AGFI) of 0.75. All of these indicators represent a poor overall fit between the hypothesized three-factor structure with 22 indicators and the observed factor pattern in the data.	Partial agreement
Convergent Validity (Construct Validity)	Multi Matrix Pourcentage of Shared Variance	r ≥ 0.40 is acceptable	Brookings, 1985	Object of comparison : Staff burnout Scale for health professional (SBS) - one factor r EE = 0.71 r PA = (-) 0.34 r DP = NR Maslach and Jackson (1981) sample r EE = 0.94 r PA = 0.94 r DP = 0.74	Total agreement

Maslach, 1981	Object of comparison: Co-worker Assessment for emotionally drained by the job and EE: r = 0.41, p < 0.01 Co-worker Assessment for emotionally drained by the job and DP: r = 0.57, p < 0.001 Co-worker Assessment for Physical fatigue and EE (Frequency): r = 0.42, p < 0.01 Co-worker Assessment for Physical fatigue and DP: r = 0.50, p < 0.01 Co-worker ratings - ""Complaints about clients "" and DP: r = 0.33, p < 0.05 Co-worker ratings of individual's satisfaction with the job and PA: r = NR Co-worker Assessment "Breaks Frequency" (Intensity) (EE): r = 0.29, p < 0.04 Co-worker Assessment "absenteeism" (DP): r = 0.30 p < 0.04 Co-worker Assessment "JDS" (EE): r = (-) 0.19, p < 0.01 Co-worker Assessment "JDS" (PA): r = 0.32, p < 0.001 Co-worker Assessment "JDS" (DP): r = (-) 0.36, p < 0.001	Disagreement
Meier, 1984	Object of comparison: Meier Burnout Assessment (MBA), r = 0.61 Self Rating of Burnout, r = 0.65 Burnout True-False, r = 0.63	Total agreement

Discriminant Validity (Construct Validity)	Multi Matrix Canonical Correlation Heterotrait-monotrait Ratio Matrix (HTMT)	r between -1 to -0.5: strong negative correlation r between -0.5 to 0: weak negative correlation r between 0 to 0.5: weak positive correlation r between 0.5 to 1: strong positive correlation A HTMT >0.80 means a lack of discriminant validity (some authors put the threshold at 0.90)	Boles, 2000	For educators' sample Meier Burnout Assessment (MBA), r = 0.61 Self-rating of Burnout, r = 0.65 Burnout True-False, r = 0.63 Among factors in the three correlated first-order factor model , r = [0.10-0.71] Parameter Estimation < 1, CI : 95% First order Three Factor Model and Two factor model (EE=DP), X2diff(2) = 108.30, p < .001) Sample Business Owners Among factors in the three correlated firt-order factor model, r = [0.07-0.71], parameter estimation < 1, CI : 95% Object of comparison : PA - EE First order Three Factor Model and Two factor model (EE=DP), X2diff(2) = 49.82, p < 0.001	Partial agreement
			Maslach, 1981	Object of comparison : JDS - General Job dissatisfaction (PA) - Frequency only, r = 0.17, p < 0.06, % of variance: < 6% " JDS - General Job dissatisfaction (DP) - Frequency only, r = - 0.22, p < 0.02, % of variance : < 6% " JDS - General Job dissatisfaction (EE), r = - 0.23, p < 0.05, % of variance : < 6%	Total agreement
			Meier, 1984	Number of comparison : 12 Number of results which met the criterion : 11 Criterion Excluded (r) : (MBA-CDD) 0.65 Object of comparison / Criterion : Validity coefficient Number of comparison : 12 Number of results which met the criterion : 10 Criterion Excluded (r) : MBA-MMPI-D (0.69) MBA-BO Self rating (0.65)" Object of comparison / Criterion : rank order of correlations within the mono-method and heterotrait-heteromethod triangles Number of triangles in the matrix: 9 Number of identical ranking: 6	Total agreement

	Reliability - Test-Retest (stability)	Fidelitiy Coefficient Structural equation modelling	1) Values >.7 are satisfactory A stable short term (2-3 weeks) dimension should have a fidelity coefficient from .8 to .9 -1 to -0.5: strong negative correlation -0.5 to 0: weak negative correlation 0 to 0.5: weak positive correlation 0.5 to 1: strong positive correlation	Maslach, 1981	Interval 2-4 Weeks. Values ranging from 0.53 to 0.82	Total agreement
Copenhagen Burnout Inventory (CBI)	Predictive Validity (Criterion Related Validity)		1) Correlation coefficient (r) r ≥.7 : good correlation, r = 0 : no correlation 2) Mixed effect regression: (β) mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. Its interpretation depends on the nature of the variables, e.g. continous or categorical.	Kristensen, 2005	Client-Related Burnout Objects of comparison: Sickness days: lowest quartile (6.9), highest quartile (13.0) Sickness spells: lowest quartile (1.5), highest quartile (2.4) Sleep problems: lowest quartile: (25.1), highest quartile: (44.6) Use of painkillers lowest quartile: 18% highest quartile: (38%) Intention to quit the workplace lowest quartile (45%), highest quartile (65%)	Total agreement

Concurrent Validity (Criterion Related Validity)		1) Correlation coefficient (r) $r \geq .7$: good correlation, $r = 0$: no correlation 2) Mixed effect regression: (β) mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. Its interpretation depends on the nature of the variables, e.g. continous or categorical.	Kristensen, 2005	Client-Related Burnout Object of comparison: Job Satisfaction: lowest quartile (68.4), highest quartile (55.1) Percentage who would choose the same job again: lowest quartile (81%), highest quartile (66%)	Total agreement
Reliability (Homogeneity)	Alpha Cronbach , α	$\alpha > 0.9$, excellent α= 0.8-0.9 & the number of itmes $\leq$10, good α= 0.8-0.9 & the number of itmes 11-30, just acceptable α= 0.7-0.8 & the number of itmes $\leq$10, acceptable α=0.6-0.7, questionable α=0.5-0.6, poor α=<0.5, unacceptable	Kristensen, 2005	Personal Burnout : 0.87 Work-Related Burnout : 0.87 Client-Related Burnout : 0.85	Total agreement
Convergent Validity (Construct Validity)	Multi Matrix Pourcentage of Shared Variance	$r \geq 0.40$ is acceptable	Kristensen, 2005	Analyses show substantial associations with job satisfaction at baseline and with sickness absence, sleep problems, use of medicine, and intention to quit three years later. The strong association between burnout and sleep problems is particularly noteworthy since fatigue/burnout and poor sleep have been shown to predict cardiovascular diseases and mortality (Prescott, et al., 2003; van Amelsvoort, Kant, Bu¨ltmann, & Swaen, 2003).	Disagreement

Discriminant Validity (Construct Validity)	Multi Matrix Canonical Correlation Heterotrait-monotrait Ratio Matrix (HTMT)	r between -1 to -0.5: strong negative correlation r between -0.5 to 0: weak negative correlation r between 0 to 0.5: weak positive correlation r between 0.5 to 1: strong positive correlation A HTMT >0.80 means a lack of discriminant validity (some authors put the threshold at 0.90)	Kristensen, 2005	The lowest correlation (divergent validity) between general health and client-related burnout.	Disagreement

Oldenburg Burnout Inventory (OLBI)	Contruct/Content Validity (Factorial analysis)	Confirmatory Analysis : GFI, RMSR, NFI, CFI, IFI, X2	Values ≥ 0.90 was considered to indicate acceptable model fit Values ≥ 0.95 is presently accepted as an indicator of good fit	Demerouti, 2001	Sample : Human resoursces GFI: 0.91, NFI : 0.84, CFI : 0.94, IFI: 0.94, RMSR: 0.05, X2 (df): 106.17 (73) Sample: Industry GFI: 0.91, NFI : 0.88, CFI : 0.97, IFI : 0.97, RMSR: 0.05, X2 (df): 194.39(73) Sample: Transport GFI: 0.90, NFI : 0.79, CFI : 0.96, IFI : 0.97, RMSR: 0.04, X2 (df): 83.06(73) Sample: Mixed GFI: 0.90 Equal factor laodings : 0.89 Equal factor variances : 0.90 Equal error vairances :0.87 NFI : 0.84 Equal factor laodings : 0.82 Equal factor variances : 0.83 Equal error vairances : 0.78 CFI : 0.95 Equal factor laodings : 0.94 Equal factor variances : 0.94 Equal error vairances : 0.9 IFI : 0.95 Equal factor laodings : 0.94 Equal factor variances : 0.94 Equal error vairances : 0.9 RMSR: 0.04 X2 (df): 303.52(219) Equal factor laodings (df): 340.52(245) Equal factor variances (df): 313.19(223) Equal error vairances (df): 409.12(250)	Partial agreement
Psychologists Burnout Inventory (PBI)	Content validity (factorial analysis)			Ackerley, 1988	1). Scree test: indicated that all four factors should be retained 2) Varimax rotation Four eigenvalues exceeding 1 (2.69, 1.93, 1.70, and 1.28), which accounted for 18%, 13%, 11%, and 9% of the variance, respectively.	Partial agreement

Burnout measure (BM)	Criterion-related validity (concurrence validity)	r correlation coefficient / concordance btwn test results and value of other variables	r ≥0.7 : good correlation r = 0 : no correlation	Pines, 1981	Variable of comparison: Satisfaction from Work Sample 1: -0.39, sample 2:- 0.53, sample 3:- 0.63, sample 4:- 0.38, sample 5:- 0.58, sample 6: -0.52, Sample 9: -0.58, sample 10 : -0.45, sample 11 : -0.37 (not significant), sample 12 : -0.45, sample 18 :- 0.53, sample 24 : -0.24, sample 25 : -0.3 - And below Israel sample, sample 26 : -0-53, sample 27 : -0.26, and sample 27 : -0.39 Satisfaction from Life: Sample 1: -0.56, sample 2:- 0.58, sample 3 :- 0.62, sample 4 :- 0.38, sample 9 : -0.44, sample 10 : -0.43, sample 11 : -0.53, sample 12 : -0.55, sample 18 :- 0.7, sample 24 : -0.34, sample 25 : -0.46 - And below Israel sample, sample 26 : 0.47, sample 27 : -0.32, and sample 28 : -0.54 Satisfaction from self: Sample 2: - 0.54, sample 3:- 0.62, sample 9: -0.45, sample 10: -0.43, sample 11: -0.34 (not significant), sample 12: -0.59, sample 18:- 0.68, sample 24 : -0.40, sample 25 : -0.41 - And below Israel sample, and sample 28: -0.32 Perception of physical health: Sample 1: 0.39, sample 4: -0.33, sample 10: -0.26, sample 24: -0.2, sample 25: -0.38 - And below Israel sample, sample 26: -0.46, sample 28: -0.28 and sample 29: -0.25 Perception for sleep problems: Sample 4: 0.30, sample 5: 0.33, and sample 6: 0.32 Conflict life and work, sample 1: 0.36, sample 10: 0.33, sample 24: 0.38, sample 29: 0.28 and sample 29: 0.24 Hopelessness (questionnaire of Beck and co): Sample 3: 0.59, p<.001" Tardiness (number of days in a year in which employees late for work): Sample 26: 0.30, p<.001" Major life events (physical and mental health, economic situation, family condition, work and other situations): Sample 3 (other sample were not assessed):	Total agreement

Positive life events	Negative life events
-0.22, p<.001	0.30, p<.001

Tendency to leave the job : Sample 5: 0.58, p< 0.5, sample 6 : 0.40, p< 0.5, sample 10 : 0.33, p< 0.5 and sample 24 : 0.27, p< 0.5

EE: emotional exhaustion, PA: personal accomplishment, DP: depersonalization, NA: not assessed, EFA: exploratory factor analysis, CFA: confirmatory factor analysis, GFI: goodness of fit index, AGFI: adjusted goodness of fit index, CFI: comparative fit index, RMSEA: root mean square error of approximation, RMR: root-mean-square residual, JD-R: job demands-resources model, X2: minimum fit function test, df: degrees of freedom, CI: confidence interval, H2: total amount of variance a variable shares with all factors, and HTMT: Heterotrait-monotrait Ratio Matrix

Supplementary table S3. Detailed results for quality assessment of five burnout PROMs according to COSMIN

PROM	Psychometric property	Overall rating	Reason for rating	Quality of evidence	Reason for downgrading the evidence
Copenhagen Burnout Inventory (CBI)	Content validity	+	The assessment of this psychometric property of the PROM design was very good but there was only one content validity study.	Moderate	We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate content validity study.
	Relevance	+	*The assessment of this psychometric property of the PROM design was very good but there was only one content validity study.*	*Moderate*	*We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate content validity study.*
	Comprehensiveness	*NA*	*The authors did not assess the comprehensiveness of their PROM.*	*Not assessed*	*We could not assess the risk of bias therefor we did not assess the quality of evidence.*
	Comprehensibility	+	*The assessment of this psychometric property of the PROM design was very good but there was only one content validity study.*	*Moderate*	*We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate content validity study.*

	Structural validity	-	No factor analysis was performed because the author thought that this validation for CBI dimensions would not be meaningful.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Internal consistency	+	They calculated the Cronbach's alpha and the indices were good.	Moderate	We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate validity study.
	Cross-cultural validity\measurement invariance	NA	We did not assess the measurement invariance in this systematic review and it was beyond our scope.	Not assessed	We did not assess the risk of bias therefor we did not assess the quality of evidence.
	Reliability	-	The authors did not perform any analysis for reliability; it was a follow-up study with three years interval.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Measurement error	-	The authors did not perform any analysis for measurement error.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Criterion validity	-	The authors did not perform any analysis for criterion validity.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Construct validity	+	The author measured the convergent and discriminant validity of their PROM, and the statistical analysis was adequate.	Moderate	We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate validity study.
	Responsiveness	-	The authors did not perform any analysis for responsiveness.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Oldenburg Burnout Inventory (OLBI)	Content validity	+	The assessment of this psychometric property of the PROM design was very good but there was only one content validity study.	Moderate/low	We downgraded the evidence to moderate-low due to indirectness of the assessment, based on comparisons between extremely different groups.

Relevance	+	*The assessment of this psychometric property of the PROM design was very good but there was only one content validity study.*	*Moderate/low*	*We downgraded the evidence to moderate-low due to indirectness of the assessment, based on comparisons between extremely different groups.*
Comprehensiveness	*NA*	*The authors did not assess the comprehensiveness of their PROM.*	*Not assessed*	*We could not assess the risk of bias therefor we did not assess the quality of evidence.*
Comprehensibility	+	*The assessment of this psychometric property of the PROM design was very good but there was only one content validity study.*	*Moderate/low*	*We downgraded the evidence to moderate-low due to indirectness of the assessment, based on comparisons between extremely different groups.*
Structural validity	+	The authors performed confirmatory factor analysis and the sample size was adequate.	Moderate/low	We downgraded the evidence to moderate-low due to indirectness of the assessment, based on comparisons between extremely different groups.
Internal consistency	+	They calculated the Cronbach’s alpha and the indices were good.	Moderate	We downgraded the evidence to moderate-low due to indirectness of the assessment, based on comparisons between extremely different groups.
Cross-cultural validity\measurement invariance	*NA*	We did not assess the measurement invariance in this systematic review and it was beyond our scope.	Not assessed	We did not assess the risk of bias therefor we did not assess the quality of evidence.
Reliability	-	The authors did not perform adequate analysis for reliability; only interrater reliabilities were estimated.	Very Low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Measurement error	-	The authors did not perform any analysis for measurement error.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Criterion validity	-	The authors did not perform any analysis for criterion validity.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Construct validity	+	The interrater reliabilities were estimated via intra-class correlation coefficients, and the statistical analysis was adequate.	Moderate	We downgraded the evidence to moderate-low due to indirectness of the assessment, based on comparisons between extremely different groups.

	Responsiveness	-	The authors did not perform any analysis for responsiveness.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Maslach Burnout Inventory (MBI)	Content validity	±	The quality of this psychometric assessment was doubtful for the PROM design and content validity studies.	Very low	We downgraded the evidence to very low due to inconsistencies in the reported results among content validity studies.
	Relevance	-	*The quality of this psychometric assessment was inadequate because the authors did not assess the relevance of the PROM.*	*Very low*	*We downgraded the evidence to very low due to high risk of bias.*
	Comprehensiveness	±	*The quality of this psychometric assessment was doubtful for the PROM design and content validity studies.*	*Very low*	*We downgraded the evidence to very low due to inconsistencies in the reported results among content validity studies.*
	Comprehensibility	±	*The quality of this psychometric assessment was doubtful for the PROM design and content validity studies.*	*Very low*	*We downgraded the evidence to very low due to inconsistencies in the reported results among content validity studies.*
	Structural validity	+	Factor analysis was conducted in many validation studies and the sample size was adequate.	Moderate	We downgraded the evidence from high to moderate because of potential risk of bias.
	Internal consistency	+	Cronbach's alpha was calculated in many validation studies and the indices were good.	High	The evidence was high and there is a low potential risk of bias.
	Cross-cultural validity\measurement invariance	NA	We did not assess the measurement invariance in this systematic review and it was beyond our scope.	Not assessed	We did not assess the risk of bias therefor we did not assess the quality of evidence.
	Reliability	-	Reliability was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Measurement error	-	Measurement error was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.

	Criterion validity	-	Criterion validity was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Construct validity	-	Inadequate analysis was conducted for construct validity.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Responsiveness	-	Responsiveness was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Burnout measure (BM)	Content validity	-	The quality of this psychometric assessment was inadequate for the PROM design and content validity studies.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Relevance	-	*The quality of this psychometric assessment was inadequate because the authors did not assess the relevance of the PROM.*	*Very low*	*There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.*
	Comprehensiveness	-	*The quality of this psychometric assessment was inadequate because the authors did not assess the comprehensiveness of the PROM.*	*Very low*	*There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.*
	Comprehensibility	-	*The quality of this psychometric assessment was inadequate for the PROM design and content validity study*	*Very low*	*There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.*
	Structural validity	-	No factor analysis was performed and the PROM was unidimensional.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Internal consistency	+	Cronbach's alpha was calculated in many validation studies and the indices were good.	Moderate	We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate validity study.

	Cross-cultural validity\measurement invariance	NA	We did not assess the measurement invariance in this systematic review and it was beyond our scope.	Not assessed	We did not assess the risk of bias therefor we did not assess the quality of evidence.
	Reliability	+	Test-retest reliability of the PROM was performed and the statistical analysis was adequate.	Moderate	We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate validity study.
	Measurement error	-	Measurement error was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Criterion validity	-	Criterion validity was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Construct validity	-	Inadequate analysis was conducted for construct validity.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Responsiveness	-	Responsiveness was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Psychologists Burnout Inventory (PBI)	Content validity	-	The quality of this psychometric assessment was inadequate for the PROM design and content validity studies.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
	Relevance	-	*The quality of this psychometric assessment was inadequate because the authors did not assess the relevance of the PROM.*	*Very low*	*There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.*
	Comprehensiveness	-	*The quality of this psychometric assessment was inadequate because the authors did not assess the comprehensiveness of the PROM.*	*Very low*	*There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.*
	Comprehensibility	-	*The quality of this psychometric assessment was inadequate for the*	*Very low*	*There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.*

		PROM design and content validity study.		
Structural validity		Principal-components factor analysis with varimax rotation were conducted.	Moderate	We downgraded the evidence from high to moderate because of potential risk of bias, as the conclusion is drawn from one adequate validity study.
Internal consistency		Internal consistency was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Cross-cultural validity\measurement invariance	?	We did not assess the measurement invariance in this systematic review and it was beyond our scope.	Not assessed	We did not assess the risk of bias therefor we did not assess the quality of evidence.
Reliability	-	Reliability was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Measurement error	-	Measurement error was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Criterion validity	-	Criterion validity was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Construct validity	-	Inadequate analysis was conducted for construct validity.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.
Responsiveness	-	Responsiveness was not tested for this PROM.	Very low	There is a high risk of bias and the quality of evidence is very low because of inadequate analysis.

Note: ±: The psychometric property assessment was inconsistent, +: The psychometric property assessment was sufficient, -: The psychometric property assessment was insufficient, and NA: The psychometric property was not assessed