

Original Article

Cite this article: Shoman Y, Marca SC, Bianchi R, Godderis L, van der Molen HF, Guseva Canu I (2021). Psychometric properties of burnout measures: a systematic review. *Epidemiology and Psychiatric Sciences* **30**, e8, 1–9. <https://doi.org/10.1017/S2045796020001134>

Received: 8 July 2020

Revised: 17 November 2020

Accepted: 7 December 2020


Key words:

Occupational Burnout; Patient-Reported Outcome Measures (PROM); psychometric properties; validity

Author for correspondence:

Yara Shoman, E-mail: yara.shoman@unisante.ch

Psychometric properties of burnout measures: a systematic review

Y. Shoman¹ , S. C. Marca¹, R. Bianchi², L. Godderis^{3,4}, H. F. van der Molen⁵ and I. Guseva Canu¹

¹Center of Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland; ²Institute of Work and Organizational Psychology, University of Neuchâtel, Neuchâtel, Switzerland; ³Department of Primary Care and Public Health, University of Leuven, Leuven, Belgium; ⁴IDEWE, External Service for Prevention and Protection at Work, Heverlee, Belgium and ⁵Amsterdam UMC, University of Amsterdam, Department of Public and Occupational Health, Netherlands Center for Occupational Diseases, Amsterdam Public Health Research Institute, Meibergdreef 9, Amsterdam, the Netherlands

Abstract

Aims. Occupational Burnout (OB) is currently measured through several Patient-Reported Outcome Measures (PROMs) and some of them have become widely used in occupational health research and practice. We, therefore, aimed to review and grade the psychometric validity of the five OB PROMs considered as valid for OB measure in mental health professionals (the Maslach Burnout Inventory (MBI), the Pines' Burnout Measure (BM), the Psychologist Burnout Inventory (PBI), the Oldenburg Burnout Inventory (OLBI) and the Copenhagen Burnout Inventory (CBI)).

Methods. We conducted systematic literature searches in MEDLINE, PsycINFO and EMBASE databases. We reviewed studies published between January 1980 and September 2018 following a methodological framework, in which each step of PROM validation, the reference method, analytical techniques and result interpretation criteria were assessed. Using the Consensus-based Standards for the selection of health Measurement Instruments we evaluated the risk of bias in studies assessing content and criterion validity, structural validity, internal consistency, reliability, measurement error, hypotheses testing and responsiveness of each PROM. Finally, we assessed the level of evidence for the validity of each PROM using the GRADE approach.

Results. We identified 6541 studies, 19 of which were included for review. Fifteen studies dealt with MBI whereas BM, PBI, OLBI and CBI were each examined in only one study. OLBI had the most complete validation, followed by CBI, MBI, BM and PBI, respectively. When examining the result interpretation correctness, the strongest disagreement was observed for MBI (27% of results), BM (25%) and CBI (17%). There was no disagreement regarding PBI and OLBI. For OLBI and CBI, the quality of evidence for sufficient content validity, the crucial psychometric property, was moderate; for MBI, BM and PBI, it was very low.

Conclusion. To be validly and reliably used in medical research and practice, PROM should exhibit robust psychometric properties. Among the five PROMs reviewed, CBI and, to a lesser extent, OLBI meet this prerequisite. The cross-cultural validity of these PROMs was beyond the scope of our work and should be addressed in the future. Moreover, the development of a diagnostic standard for OB would be helpful to assess the sensitivity and specificity of the PROMs and further reexamine their validity.

The study protocol was registered in PROSPERO (CRD 42019124621).

© The Author(s), 2021. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (<http://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.

Introduction

Occupational Burnout (OB) is a relatively recent entity that was first mentioned in the literature in the late 1960s (Bradley, 1969). The following 50 years of uncoordinated research resulted in multiple, somehow-contradictory definitions and measures of OB worldwide. The current situation reflects this semantic and methodological heterogeneity: even the World Health Organization (WHO) is uncertain how to deal with OB. The WHO included burnout in the tenth revision of the international classification of diseases, but in the forthcoming eleventh revision (WHO, 2019), specified that it was a phenomenon and not a disease.

Nowadays, the application of the Evidence-Based Medicine (EBM) in diagnostic and prognostic processes used in healthcare is essential (Newman and Kohn, 2009). However, the lack of harmonisation regarding acceptable validity standards or criteria for various mental health measures (Haberer *et al.*, 2013) directly challenges the EBM application in diagnosis and, subsequently, in treatment of mental health disorders. With respect to OB, this lack of harmonisation in OB definition and measure is particularly salient, precluding a reliable estimation of its prevalence (Rotenstein *et al.*, 2018) and triggering to exaggeration of this phenomenon as a

21st century epidemic (Bianchi, 2017; Mirkovic and Bianchi, 2019) or sometimes it can result in underestimation (Doulougeri *et al.*, 2016). Therefore, the Network on the Coordination and Harmonization of European Occupational Cohorts (OMEGA-NET) decided to prioritise this issue (Guseva Canu *et al.*, 2019) and to propose a harmonised definition of OB as a health outcome to be used in future longitudinal studies (Guseva Canu *et al.*, 2020). The next step is thus to harmonise the measurement of OB.

There is no consensus on the measurement of OB (Poghosyan *et al.*, 2009) and all identified published measures are Patient Reported Outcome Measures (PROMs) (Rotenstein *et al.*, 2018; Guseva Canu *et al.*, 2020), i.e., measures completed by the patient (Jokstad, 2018). There are about a dozen different OB PROMs, eight of which were considered as valid for measuring OB in mental health professionals (O'Connor *et al.*, 2018), including the Maslach Burnout Inventory (MBI) (Maslach and Jackson, 1981), the Pines' Burnout Measure (BM) (Malakh-Pines *et al.*, 1981), the Psychologist Burnout Inventory (PBI) (Ackerley *et al.*, 1988), the Oldenburg Burnout Inventory (OLBI) (Demerouti *et al.*, 2001), the Professional Quality of Life Measure (ProQOL) (Stamm, 2010), the Copenhagen Burnout Inventory (CBI) (Kristensen *et al.*, 2005), the Children Services Survey (CSS) (Glisson and Hemmelgarn, 1998) and the Organizational Social Context (OCS) (Glisson *et al.*, 2008). Considering the diversity of these PROMs, a closer look at their validity should inform their use in medical research and practice. The objectives of this systematic review were to assess the validation processes used in each of the selected PROMs and to grade the evidence of psychometric quality to recommend the most valid PROM(s) for use in medical practice and epidemiological research on OB.

Methods and analysis

We performed this systematic review following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (Moher *et al.*, 2009).

Protocol and registration

A review protocol is available on the international database PROSPERO with the registration number CRD42019124621 on the following link: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=124621.

Eligibility criteria

We searched for studies assessing the psychometric properties of eight OB PROMs considered as validated for measuring burnout in mental health professionals (O'Connor *et al.*, 2018): MBI, BM, PBI, OLBI, ProQOL, CBI, CSS and OCS. Henceforth we focussed on PROMs dealing with OB exclusively, leading to the final inclusion of five PROMs. We excluded ProQOL, CSS and OCS because ProQOL measures burnout as a dimension, at the same level as the secondary trauma (Stamm, 2010), while CSS and OCS measure the organisational aspects influencing services efficiency (Glisson and Hemmelgarn, 1998; Glisson *et al.*, 2008). We included studies (1) with quantitative testing of psychometric properties; (2) published as original research articles; (3) addressing psychometric properties of at least one included OB PROMs in its original (not translated) version; (4) with a sample size of

>100 participants. We excluded studies (1) for which no full text could be found; (2) where one of the five burnout PROMs was used as a reference against another one, not included in this review; (3) where participants were not professionally employed (e.g., students, medical residents).

Data sources and search terms

We performed a systematic literature search for the period from 01/01/1980 to 27/09/2018. This time window was defined based on the fact that the first OB PROM, MBI, dates from 1981. We used three databases to search for eligible studies via the online catalogue of databases OVID interface: MEDLINE, PsycINFO and EMBASE. An experienced librarian reviewed the search strategy that consisted of free-text words to specify three search strings: terms focusing on the burnout PROM of interest (e.g., MBI), terms related to the validation of the PROM and a combination of the two first search strings results. Finally, one additional search string consisted of removing duplicates. In addition, we checked the reference lists from articles and reviews retrieved in our electronic search for any additional studies to include. For the PROM for which no article was found, we searched for their primary sources (e.g., books), and included them in this review. The full search strategy is available in online Supplementary Table S1.

Data collection and analysis

Study selection

We imported the collected studies in the bibliography software EndNote X8 and selected the studies in a three-step process done by two independent reviewers (SCM and YS). First, the reviewers eliminated possible remaining duplicates within each database and between databases. Second, they examined the title and the abstract of each article. They retained or rejected articles based on the above-mentioned inclusion and exclusion criteria. Third, the reviewers read the full-text of the remaining articles and followed the same procedure with the selected articles. For each of the three steps, reviewers discussed all discrepancies in the assessment of the studies and, when needed, consulted a third reviewer (IGC).

Data extraction and management

We extracted the data through a two-step process. First, we developed a standardised data extraction form convenient for all kinds of study designs and methods applied. Each burnout PROM had its own exemplary data extraction form (MS Excel file). Two independent reviewers tested the form using articles on different burnout PROMs. They discussed the discrepancies and if needed, consulted a third reviewer for clarification and decision. This process continued until a complete agreement was reached between reviewers on the finalised data extraction form. Then, the two reviewers independently extracted the data and compared their results. The extracted data concerned studies' identification (i.e., authors, year of publication, journal and title); samples' characteristics (i.e., size, sex ratio, age, occupational activity, participation rate, representativity, OB scores' distribution); burnout PROMs' characteristics (i.e., name, version, number of items, number of dimensions, dimensions' names); and statistical methods used for assessing the psychometric properties outcome. We identified

the missing data by a code depending on the reason why they are missing (not assessed *v.* not reported). Secondly, we developed an additional table, in which we extracted quantitative results for each psychometric property for their further analysis.

Validity assessment and grading

We analysed the collected data in four steps. Each step was conducted independently by two reviewers and cross-checked by two other reviewers.

Validity completeness assessment

First, we counted the number of psychometric properties (i.e., face validity, content validity, predictive validity, concurrent validity, convergent validity, discriminant validity, exploratory factorial validity, confirmatory factorial validity, stability, homogeneity and sensitivity) assessed for each burnout PROM. For example, if a study analysed the psychometric property with an exploratory factorial analysis, a confirmatory factorial analysis and a coefficient of internal consistency, three psychometric properties (i.e., exploratory factorial validity, confirmatory factorial validity and internal consistency) were counted. This enabled assessing the completeness of validation for each burnout PROM considered.

Quantitative assessment of psychometric validity

Second, we examined the reported quantitative results and interpreted them using a previously established methodological framework (Marca *et al.*, 2020). This framework specifies for each psychometric property, its definition, the method recommended for its analysis, resulting statistics and objective criteria for their interpretation. To assess the correctness of conclusion on validity for each psychometric property of a PROM, we compared the result interpretation by the authors with results interpretation according to the framework. We made this comparison for each burnout dimension separately and rated the degree of discrepancy. The comparison between the interpretations of the authors and the reviewers resulted in a complete agreement when there was no discrepancy between them. A partial agreement corresponded to differences in cutoff values, e.g. a correlation of 0.50 considered as moderate in framework and the authors considered it as strong. A disagreement corresponded to an overall interpretation discrepancy, e.g. the authors interpreted a model as acceptable and the reviewers as not acceptable based on fit indices norms. No comparison was possible when the interpretation of the authors was missing.

Risk of bias assessment

We assessed the risk of bias of each PROM validation study according to the COnsensus-based Standards for the selection of health Measurement INSTRUMENTS (COSMIN) checklist (Mokkink *et al.*, 2010). COSMIN triggers rating of PROM development study and content validity studies as very good, adequate, doubtful and not assessed. It assesses the content validity of a PROM through measuring the relevance, comprehensiveness and comprehensibility. Moreover, it considers eight other psychometric properties: structural validity, internal consistency, reliability, measurement error, criterion validity, cross-cultural validity, measurement invariance, hypotheses testing and responsiveness.

Quality assessment

Finally, we graded the quality of evidence on psychometric validity of each burnout PROM following the modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Terwee *et al.*, 2018). According to the GRADE (Guyatt *et al.*, 2008), there are four levels for the quality of evidence: very low, low, moderate and high. When assessing the quality of evidence for a PROM's validity using the GRADE, the risk of bias, consistency, directness and precision of studies available for each PROM, should be considered together. We started by assuming that the quality of evidence from the studies is high, then we downgraded it depending on the risk of bias, inconsistency, indirectness and imprecision.

Results

Selected studies

The literature search resulted in 6541 references and 5442 remained after removing the duplicates (Fig. 1). Seventy-six studies were selected for the full-text screening, of which 16 were eligible; three additional studies were identified from reference lists. Overall, 19 studies were thus included in the review, 15 of which dealt with MBI (Iwanicki and Schwab, 1981; Maslach and Jackson, 1981; Gold, 1984; Meier, 1984; Brookings *et al.*, 1985; Lahoz and Mason, 1989; Gold *et al.*, 1992; Holland *et al.*, 1994; Yadama and Drake, 1995; Boles *et al.*, 2000; Kalliath and O'Driscoll, 2000; Beckstead, 2002; Kim and Ji, 2009; Poghosyan *et al.*, 2009; Chao *et al.*, 2011) whereas BM, PBI, OLBI and CBI were each examined in one study only (Table 1).

Results of completeness and quantitative assessment of psychometric validity

MBI and CBI had the most complete validation, with seven psychometric properties assessed out of 11 (Table 1). PBI had the lowest validation completeness with one psychometric property assessed, namely the factorial validity. The results of the agreement between the authors' and the reviewers' interpretations of quantitative results are reported in online Supplementary Table S2. For MBI, we found partial agreement on five analyses of psychometric properties: the discriminant validity (Boles *et al.*, 2000), factorial validity based on exploratory factor analysis (Poghosyan *et al.*, 2009) and confirmatory factor analysis (Yadama and Drake, 1995), and reliability based on Cronbach's alpha (Brookings *et al.*, 1985; Boles *et al.*, 2000). We found 11 disagreements related to the convergent validity of MBI (Maslach and Jackson, 1981), factorial validity based on exploratory (Lahoz and Mason, 1989; Holland *et al.*, 1994) and confirmatory (Gold, 1984; Gold *et al.*, 1992; Holland *et al.*, 1994; Boles *et al.*, 2000; Kim and Ji, 2009; Poghosyan *et al.*, 2009) factor analyses and reliability measured via Cronbach's alpha (Meier, 1984; Kalliath and O'Driscoll, 2000). As we analysed each dimension separately, the exploratory factor analysis is shown in online Supplementary Table S2 with eigenvalues for each dimension and for intensity and frequency (online Supplementary Table S2). However, we disagreed with these results as the value of communality has to be ≥ 0.90 to indicate acceptable model fit and the reported value for frequency was 51% and intensity 50.6%. For PBI, we had one partial agreement related to the factorial validity specifically exploratory factor analysis. We had one partial agreement with OLBI concerning factorial validity specifically

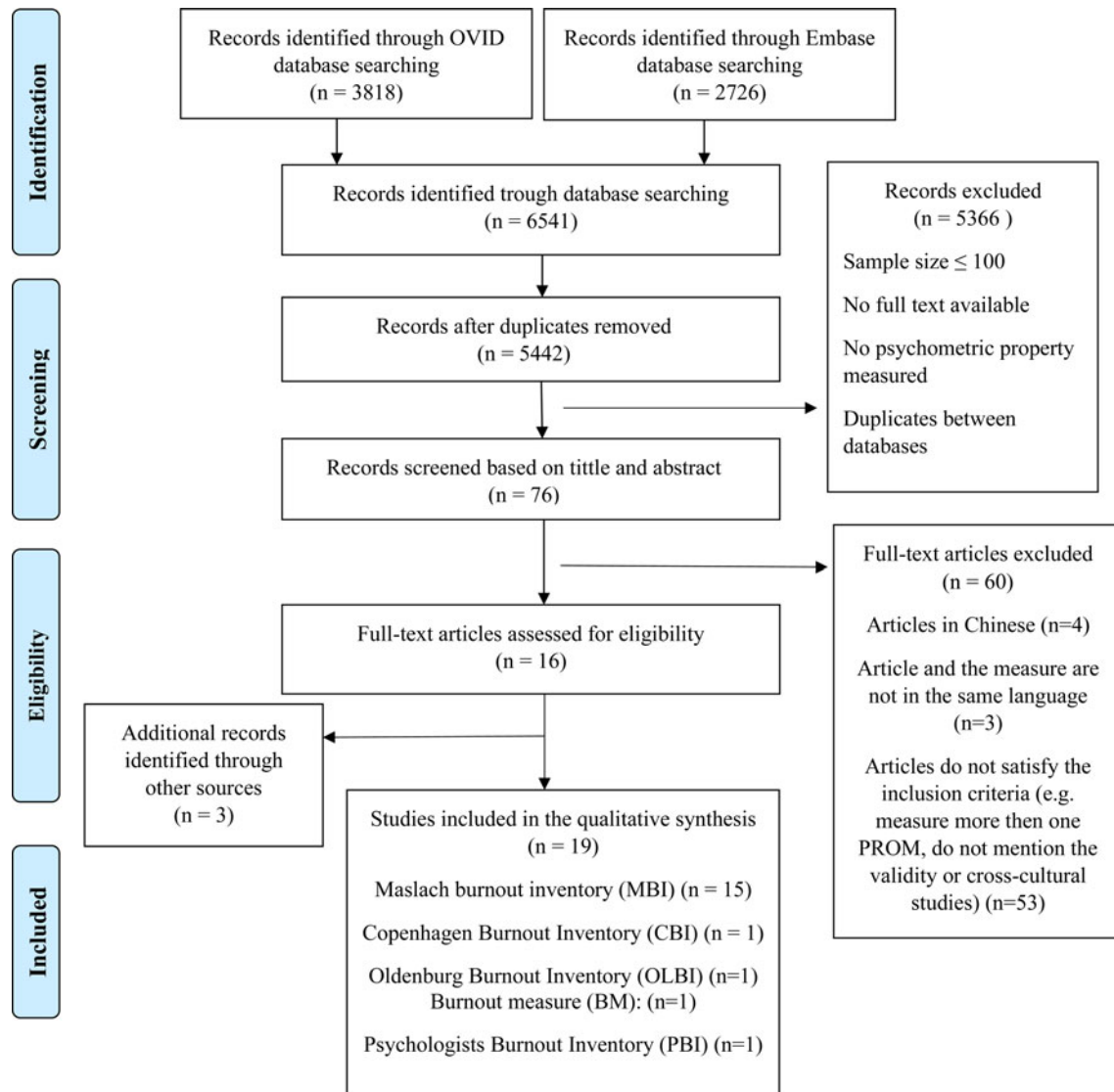


Fig. 1. Flow-chart of the included studies.

confirmatory factor analysis. For CBI, we had a disagreement related to discriminant validity. While internal consistency and factorial validity were widely assessed for most PROMs, we found no formal content validity study for any of the selected PROMs.

Results of risk of bias and quality assessment

When assessing the content validity of MBI according to COSMIN, we found no data on its relevance, whereas the results on its comprehensiveness and comprehensibility were inconsistent across studies (Table 2 and online Supplementary Table S3). For instance, some authors recommended the original MBI structure with three dimensions (Gold, 1984; Gold *et al.*, 1992; Kim and Ji, 2009) while others recommended a modified structure, limited to two dimensions (Kalliath and O'Driscoll, 2000), or a four-item reduced original structure (Yadama and Drake, 1995). We also revealed the inconsistency in rating each item of MBI for frequency, for intensity, or for both. Based on these results, we downgraded the quality of evidence for content validity of MBI from doubtful to very low (online Supplementary Table S3). For BM and PBI, the quality of evidence on content

validity was also very low, because their relevance, comprehensiveness and comprehensibility were not validated using adequate analysis. For OLBI, we downgraded the quality of content validity to moderate/low due to the indirectness of its assessment, based on comparisons between extremely different groups. The CBI achieved the highest level of evidence for content validity, although the authors did not assess its comprehensiveness. According to the COSMIN, insufficient content validity could have been a stopping point for assessing the PROM validity. Nevertheless, we considered the seven other properties from the COSMIN checklist to enable meaningful and complete comparison of PROMs. For these properties, OLBI achieved the highest grade, with three validated psychometric properties (structural validity, internal consistency and construct validity), whereas CBI, BMI and BM completed two of them and PBI only assessed its structural validity (Table 2). It appears that the structural validity and the internal consistency are the most assessed psychometric properties, while the measurement error, the known-groups validity and the responsiveness were never assessed. It is noteworthy that the absence of an accepted diagnostic standard precluded measuring sensitivity and specificity of all OB PROMs (Table 1).

Table 1. Burnout PROMs' description, initial validation performed and validity of statistical analysis interpretation

PROM	Maslach Burnout Inventory	Burnout Measure	Psychologists Burnout Inventory	OLdenburg Burnout Inventory	Copenhagen Burnout Inventory
First author	C. Maslach	A. Pines	G. D. Ackerley	E. Demerouti	T.S. Kristensen
Aim	Assess various aspects of the burnout syndrome	Describe the tedium measure	Examine the level of burnout in USA psychologists	Use a new burnout inventory through two aspects	Discuss a new questionnaire
Year of publication	1981	1981	1988	1999	2005
Country	USA	USA	USA	Germany	Denmark
Language of original version	English	English	English	German	Danish
Population	Workers with a regular contact with other people	General population	Psychologists	General population	Population working in human service sector
Dimensions (Number of items)	Emotional Exhaustion (9)	Physical exhaustion (7)	Control (3)	Exhaustion (7)	Personal burnout (6)
	Depersonalisation (5)	Emotional exhaustion (7)	Support (3)	Disengagement (8)	Work-related burnout (7)
	Personal accomplishment (8)	Mental exhaustion (7)	Negative Clientele (6)		Client-related burnout (6)
			Over involvement (3)		
Number of assessed articles	15	1	1	1	1
Number of psychometric properties assessed	7/11	4/11	1/11	4/11	7/11
Face validity	√ (1)	-	-	-	√ (1)
Content validity	-	-	-	-	-
Predictive validity	-	-	-	-	√ (1)
Concurrent validity	-	-	-	-	√ (1)
Convergent validity	√ (3)	√ (1)	-	-	√ (1)
Discriminant validity	√ (3)	√ (1)	-	-	√ (1)
Exploratory factorial validity	√ (7)	-	√ (1)	-	-
Confirmatory factorial validity	√ (11)	-	-	√ (1)	-
Stability	√ (3)	√ (1)	-	-	√ (1)
Homogeneity	√ (13)	√ (1)	-	√ (3)	√ (1)
Sensitivity	-	-	-	-	-
Comparison between authors' and reviewers' interpretation of results					
Agreement	24/41	1/4	0/1	0/4	4/7
Partial agreement	5/41	0/4	1/1	1/4	0/7
Disagreement	11/41	0/4	0/1	0/4	1/7
Impossible agreement	1/41	3/4	0/1	1/4	1/7

Discussion

Main findings

For CBI, we found moderate quality of evidence on its content validity, internal consistency and construct validity, but very low quality of evidence of structural validity, reliability, measurement

error, criterion validity and responsiveness. OLBI had a moderate to low quality of evidence for content validity, construct validity and structural validity and moderate quality of evidence for internal consistency. With this performance, OLBI had the highest number of psychometric properties assessed among the five reviewed PROMs. MBI, BM and PBI had a very low quality of

Table 2. Systematic review results for five burnout PROMs according to COSMIN

COSMIN psychometric properties	Maslach Burnout Inventory		Burnout Measure		Psychologists Burnout Inventory		Oldenburg Burnout Inventory		Copenhagen Burnout Inventory	
	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence	Overall rating	Quality of evidence
Content validity	±	Very low	-	Very low	-	Very low	+	Moderate/Low	+	Moderate
Structural validity	+	Moderate	-	Very low	+	Moderate	+	Moderate/Low	-	Very low
Internal consistency	+	High	+	Moderate	-	Very low	+	Moderate	+	Moderate
Reliability	-	Very low	+	Moderate	-	Very low	-	Very low	-	Very low
Measurement error	-	Very low	-	Very low	-	Very low	-	Very low	-	Very low
Criterion validity	-	Very low	-	Very low	-	Very low	-	Very low	-	Very low
Construct validity	-	Very low	-	Very low	-	Very low	+	Moderate/Low	+	Moderate
Responsiveness	-	Very low	-	Very low	-	Very low	-	Very low	-	Very low

Note: ±, the psychometric property assessment was inconsistent; +, the psychometric property assessment was sufficient; -, the psychometric property assessment was insufficient.

evidence for content validity. Nevertheless, the psychometric properties of MBI were the most studied among the five PROMs and most of them were interpreted correctly.

Results' interpretation

Based on the evidence assessed by an objective multi-step approach, CBI appeared the most valid of the five reviewed PROMs, but essentially because of its content validity. Nevertheless, it is important to mention that CBI validation was completed by its authors in only one study, though a very comprehensive one. Most of their results were interpreted correctly; we found a slight over-interpretation regarding only one psychometric property (discriminant validity). CBI is the most recent PROM (2005), which can justify measuring more psychometric properties than the other, older PROMs. However, as it was originally developed in Danish, the overall evidence on its validity beyond the content validity is still insufficient to recommend CBI as the best OB PROM based on this review. As CBI was translated into different languages (e.g., English, German, French, Spanish, Chinese and Korean) and utilised in several countries, where it was judged as a robust PROM for OB (Milfont *et al.*, 2008; Molinero Ruiz *et al.*, 2013; Fong *et al.*, 2014; Fiorilli *et al.*, 2015; Phuekphan *et al.*, 2016; Javanshir *et al.*, 2019; Jeon *et al.*, 2019), the cross-cultural validity of the translated versions should be assessed.

OLBI was developed to tackle some drawbacks of MBI, especially the wording of the dimensions (Demerouti *et al.*, 2001; Halbesleben and Demerouti, 2005). According to our findings, OLBI is the second most valid available PROM of OB. This rating is due to the indirectness that downgraded the quality of evidence of its content validity but a larger number of psychometric properties assessed according to COSMIN checklist compared to CBI. Compared to CBI, OLBI's validation completeness was lower according to the methodological framework. However, we found no disagreement with the interpretation of its validation. OLBI overcame the limitations of MBI by balanced wording and broader conceptualisation of burnout, which is not restricted to human service's workers (Demerouti *et al.*, 2001; Halbesleben and Demerouti, 2005). MBI has negative wording for emotional exhaustion and depersonalisation and positive wording for personal accomplishment dimension, leading to a potential wording bias. Conversely, OLBI has both positive and negative worded items (Demerouti *et al.*, 2003), is shorter than MBI and publically available in different languages. These features likely explain why OLBI is the second most used OB PROM after MBI (Guseva Canu *et al.*, 2020). BM is the oldest among the five PROMs reviewed in our study. Some studies reported that BM is reliable and valid (Pines and Aronson, 1988; Pines, 1993; Schaufeli and Van Dierendonck, 1993; Schaufeli and Enzmann, 1998). We found inadequate content validity with a very low quality of evidence of psychometric validity for BM as well as for PBI. The latter dates back to 1988 and the study that dealt with it was not focused on the psychometric analysis of the PROM but rather on its comparison with MBI (Ackerley *et al.*, 1988).

As expected, MBI validity was studied more than for other PROMs, probably because MBI remains the most used OB PROM (Guseva Canu *et al.*, 2020). Some authors considered MBI as the gold standard for OB PROMs (Maslach *et al.*, 1981; West *et al.*, 2012; Williamson *et al.*, 2018), which can be debated provided the results of this review. The subsequent development of OLBI and CBI confirms the unsatisfactory features of MBI

and the need of a diagnostic standard for OB (Arvidsson *et al.*, 2016; Rotenstein *et al.*, 2018). In MBI, emotional exhaustion is often considered separately, representing the core of burnout syndrome (Maslach and Jackson, 1981; Kristensen *et al.*, 2005) but also of depression or along with depersonalisation dimension to represent the core of burnout (Bussing and Glaser, 2000). Some authors argue that depersonalisation and personal accomplishment are not even a part of OB (Kristensen *et al.*, 2005). Concerning the overall psychometric validity, MBI has very low quality of evidence on validity for six psychometric properties out of eight although it had the highest number of validation studies.

It is worth noting that most PROMs were developed and assessed well before the methodological guidelines and frameworks for PROMs validation became available. This might partly explain the insufficient psychometric quality and validation completeness of the PROMs reviewed in this study.

Strength and limitation

This review assesses the evidence on psychometric validity of five commonly used PROMs. Besides its originality and topicality, this work has several methodological strengths, including the robustness of the research protocol, the exhaustiveness of the literature search, performed with assistance of an experienced documentarist using three important databases over a 40-year period. Every step of screening, data extraction, analysis and quality assessment was performed by two reviewers independently and double-checked by a third reviewer. For validity assessment, we used two complementary methods: our own methodological framework developed for validation of PROMs (Marca *et al.*, 2020) and the international standardised method (Mokkink *et al.*, 2010). The latter was completed with a modified-GRADE assessment after we started this study (Terwee *et al.*, 2018). While the methodological framework allows assessing the completeness of validation and facilitates the objective results interpretation, the COSMIN is helpful in assessing content validity studies, the most important psychometric property of a PROM. Therefore, using these methods together enabled us analysing all aspects of qualitative and quantitative approaches used in PROMs validation thoroughly and providing methods triangulation.

The content validity is assessed based on the PROM's development study, which implies the use of original and not translated PROM version. Therefore, we did not consider studies using translated versions of selected PROMs. After the validation of the original version, the translated version should follow the process of cross-cultural validity assessment (Beaton *et al.*, 2000; Terwee *et al.*, 2018). Correlations may differ according to countries and this emphasises the significance of cross-cultural validity (Pines *et al.*, 2002). As our results suggest moderate quality of evidence of the content validity of CBI and OLBI, their cross-cultural validity assessment is highly recommended. It is noteworthy that four different French versions of OLBI currently co-exist (Belgian, French, Canadian and Swiss).

The small number of studies included in this review is a limitation, precluding firm conclusion on the quality of evidence of the reviewed PROMs. Considering a large timespan for the systematic literature search, allowed us observing that often PROM's validity results were published either as part of the PROM development study or shortly after. Therefore, the limitation of the systematic search 27/09/2018 should not be considered problematic, given that the last PROM was published in 2005.

However, more methodologically robust validation studies are necessary for verifying results consistency and for the development of a diagnostic standard for OB.

Finally, as we only considered five OB PROMs cited as valid for assessing burnout in mental health professionals by O'Connor *et al.*, some OB PROMs, such as Shirom-Melamed Burnout Measure (SMBM), remained beyond of our assessment. A recent study by Schilling *et al.* (2019) concluded that SMBM validity and reliability were rarely examined in the literature. However, given the widespread use of SMBM, an assessment of its validity in future research is suitable.

Suggestions for future research in the field

Future research should further examine the psychometric properties that were insufficiently assessed or valid in CBI and OLBI, and assess all other available OB PROMs' validity. The development of a diagnostic standard for OB is a priority. It will facilitate OB PROMs comparison through the assessment of their sensitivity, specificity and diagnostic accuracy.

Conclusions

To be validly and reliably used in medical research and practice, PROMs should exhibit robust psychometric properties. Among the five PROMs that we reviewed (CBI, MBI, OLBI, BM and PBI), only CBI and, to a lesser extent, OLBI were able to meet this prerequisite. The cross-cultural validity of these PROMs was beyond the scope of our work and should be addressed in the future. Moreover, the development of a diagnostic standard for OB would be helpful to assess the sensitivity and specificity of the PROMs and further establish their validity.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S2045796020001134>.

Availability of data and materials. Data are available online as supplementary material of the present article.

Acknowledgements. The authors thank Aline Sager, Christina Gyorkos and Paola Paatz for their precious help in establishing the search queries and screening.

Financial support. University of Lausanne and University of Bern BNF – National Qualification Program funded the salary of young researcher (SCM). European Cooperation in Science & Technology (COST Action CA16216), OMEGA-NET: Network on the Coordination and Harmonization of European Occupational Cohorts covered the meetings and travel expenses as well as the open access publication costs. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801076, through the SSPH + Global PhD Fellowship Programme in Public Health Sciences (GlobalP3HS) of the Swiss School of Public Health.

Conflicts of interest. None.

Ethical standards. This research is a systematic review of available studies and does not involve human and/or animal experimentation. The Ethics committee's approval was not required.

References

Ackerley GD, Burnell J, Holder DC and Kurdek LA (1988) Burnout among licensed psychologists. *Professional Psychology-Research and Practice* **19**, 624–631.

- Arvidsson I, Hakansson C, Karlson B, Bjork J and Persson R (2016) Burnout among Swedish school teachers – a cross-sectional analysis. *BMC Public Health* **16**, 823.
- Beaton DE, Bombardier C, Guillemin F and Ferraz MB (2000) Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)* **25**, 3186–3191.
- Beckstead JW (2002) Confirmatory factor analysis of the Maslach Burnout Inventory among Florida nurses. *International Journal of Nursing Studies* **39**, 785–792.
- Bianchi R (2017) Is the “burnout epidemic” an academic fiction? *BMJ (Clinical research ed.)* **358**, j4389.
- Boles JS, Dean DH, Ricks JM, Short JC and Wang GP (2000) The dimensionality of the Maslach Burnout Inventory across small business owners and educators. *Journal of Vocational Behavior* **56**, 12–34.
- Bradley HB (1969) Community-based treatment for young adult offenders. *Crime & Delinquency* **15**, 359–370.
- Brookings JB, Bolton B, Brown CE and Mcevoy A (1985) Self-reported job burnout among female human-service professionals. *Journal of Occupational Behaviour* **6**, 143–150.
- Bussing A and Glaser J (2000) Four-stage process model of the core factors of burnout: the role of work stressors and work-related resources. *Work and Stress* **14**, 329–346.
- Chao SF, McCallion P and Nickle T (2011) Factorial validity and consistency of the Maslach Burnout Inventory among staff working with persons with intellectual disability and dementia. *Journal of Intellectual Disability Research* **55**, 529–536.
- Demerouti E, Bakker AB, Nachreiner F and Schaufeli WB (2001) The job demands-resources model of burnout. *Journal of Applied Psychology* **86**, 499–512.
- Demerouti E, Bakker AB, Vardakou I and Kantas A (2003) The convergent validity of two burnout instruments – a multitrait-multimethod analysis. *European Journal of Psychological Assessment* **19**, 12–23.
- Doulougeri K, Georganta K and Montgomery A (2016) “Diagnosing” burnout among healthcare professionals: can we find consensus? *Cogent Medicine* **3**, 275–281.
- Fiorilli C, De Stasio S, Beneve P, Iezzi DF, Pepe A and Albanese O (2015) Copenhagen Burnout Inventory (CBI): a validation study in an Italian teacher group. *Testing Psychometrics Methodology in Applied Psychology* **22**, 537–551.
- Fong TCT, Ho RTH and Ng SM (2014) Psychometric properties of the Copenhagen burnout inventory-Chinese version. *Journal of Psychology* **148**, 255–266.
- Glisson C and Hemmelgarn A (1998) The effects of organizational climate and interorganizational coordination on the quality and outcomes of children’s service systems. *Child Abuse & Neglect* **22**, 401–421.
- Glisson C, Landsverk J, Schoenwald S, Kelleher K, Hoagwood KE, Mayberg S, Green P and Health RNOYM (2008) Assessing the organizational social context (OSC) of mental health services: implications for research and practice. *Administration and Policy in Mental Health and Mental Health Services Research* **35**, 98.
- Gold Y (1984) The factorial validity of the Maslach burnout inventory in a sample of California elementary and junior-high school classroom teachers. *Educational and Psychological Measurement* **44**, 1009–1016.
- Gold Y, Roth RA, Wright CR, Michael WB and Chen CY (1992) The factorial validity of a teacher burnout measure (educators survey) administered to a sample of beginning teachers in elementary and secondary-schools in California. *Educational and Psychological Measurement* **52**, 761–768.
- Guseva Canu I, Mesot O, Gyorkos C, Mediouni Z, Mehlum IS and Bugge MD (2019) Burnout syndrome in Europe: towards a harmonized approach in occupational health practice and research. *Industrial Health* **57**, 745–752.
- Guseva Canu I, Marca SC, Dell’oro F, Balázs Á, Bergamaschi E, Besse C, Bianchi R, Bislimovska J, Bjelajac AK, Buggez M, Busneag CL, Çağlayan C, Cernițanu M, Pereira CC, Hafner ND, Droz N, Eglite M, Godderis L, Gündel H, Hakanen JJ, Iordache RM, Khireddine-Medouni I, Kiran S, Larese-Filon F, Lazor-Blanchet C, Légeron P, Loney T, Majery N, Merisalu E, Mehlum IS, Michaud L, Mijakoski D, Minov J, Modenese A, Molan M, Van Der Molen HF, Nena E, Nolimal D, Otelea M, Pletea E, Pranjic N, Rebergen D, Reste J, Schernhammer E and Wahlen A (2020) Harmonized burnout definition, finally. A systematic review, semantic analysis, and Delphi consensus in 29 countries. *Scandinavian Journal of Work, Environment & Health*. doi: 10.5271/sjweh.3935
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ and Group GW (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal* **336**, 924–926.
- Haberer JE, Trabin T and Klinkman M (2013) Furthering the reliable and valid measurement of mental health screening, diagnoses, treatment and outcomes through health information technology. *General Hospital Psychiatry* **35**, 349–353.
- Halbesleben JRB and Demerouti E (2005) The construct validity of an alternative measure of burnout: investigating the English translation of the Oldenburg Burnout Inventory. *Work and Stress* **19**, 208–220.
- Holland PJ, Michael WB and Kim S (1994) Construct-validity of the educators survey for a sample of middle school teachers. *Educational and Psychological Measurement* **54**, 822–829.
- Iwanicki EF and Schwab RL (1981) A cross validation-study of the Maslach burnout inventory. *Educational and Psychological Measurement* **41**, 1167–1174.
- Javanshir E, Dianat I and Asghari-Jafarabadi M (2019) Psychometric properties of the Iranian version of the Copenhagen Burnout Inventory. *Health Promotion Perspectives* **9**, 137–142.
- Jeon GS, You SJ, Kim MG, Kim YM and Cho SI (2019) Psychometric properties of the Korean version of the Copenhagen burnout inventory in Korean homecare workers for older adults. *PLoS ONE* **14**, e0221323.
- Jokstad A (2018) Patient-reported outcomes (PROs) versus patient-reported outcome measures (PROMs) Is there a difference? *Clinical and Experimental Dental Research* **4**, 61–62.
- Kalliath TJ and O’Driscoll MP (2000) A test of the Maslach Burnout Inventory in three samples of healthcare professionals. *Work and Stress* **14**, 35–50.
- Kim H and Ji JY (2009) Factor structure and longitudinal invariance of the Maslach burnout inventory. *Research on Social Work Practice* **19**, 325–339.
- Kristensen TS, Borritz M, Villadsen E and Christensen KB (2005) The Copenhagen burnout inventory: a new tool for the assessment of burnout. *Work and Stress* **19**, 192–207.
- Lahoz MR and Mason HL (1989) Maslach burnout inventory – factor structures and norms for USA pharmacists. *Psychological Reports* **64**, 1059–1063.
- Malakh-Pines A, Aronson E and Kafry D (1981) *Burnout: From Tedium to Personal Growth*. USA: Free Press.
- Marca SC, Paatz P, Gyorkos C, Cuneo F, Bugge MD, Godderis L, Bianchi R and Guseva Canu I (2020) Validation of questionnaires and rating scales used in medicine: protocol for a systematic review of burnout self-reported measures. *medRxiv*. <https://doi.org/10.1101/2020.06.24.20138115>.
- Maslach C and Jackson SE (1981) The measurement of experienced burnout. *Journal of Occupational Behaviour* **2**, 99–113.
- Maslach C, Jackson SE and Leiter MP (1981) Maslach burnout inventory. *Journal of Occupational Behaviour* **2**, 99–113.
- Meier ST (1984) The construct-validity of burnout. *Journal of Occupational Psychology* **57**, 211–219.
- Milfont TL, Denny S, Ameratunga S, Robinson E and Merry S (2008) Burnout and wellbeing: testing the Copenhagen burnout inventory in New Zealand teachers. *Social Indicators Research* **89**, 169–177.
- Mirkovic D and Bianchi R (2019) Physician burnout: let’s avoid unsubstantiated claims. *Nature Reviews Clinical Oncology* **16**, 136–136.
- Moher D, Liberati A, Tetzlaff J and Altman DG and PRISMA Group (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* **6**, e1000097.
- Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM and De Vet HC (2010) The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research* **19**, 539–549.
- Molinero Ruiz E, Basart Gomez-Quintero H and Moncada Lluís S (2013) [Validation of the Copenhagen burnout inventory to assess professional burnout in Spain]. *Revista Española de Salud Pública* **87**, 165–179.
- Newman TB and Kohn MA (2009) *Evidence-Based Diagnosis*. United States of America: Cambridge University Press.

- O'Connor K, Muller Neff D and Pitman S** (2018) Burnout in mental health professionals: a systematic review and meta-analysis of prevalence and determinants. *European Psychiatry* **53**, 74–99.
- Phuekphan P, Aungsuroch Y, Yunibhand J and Chan SWC** (2016) Psychometric properties of the Thai version of Copenhagen Burnout Inventory (T-Cbi) in Thai nurses. *Journal of Health Research* **30**, 135–142.
- Pines AM** (1993) *Burnout: An Existential Perspective*. Philadelphia, PA, USA: Taylor & Francis.
- Pines AM and Aronson E** (1988) *Career Burnout*. New York: Free Press.
- Pines AM, Ben-Ari A, Utasi A and Larson D** (2002) A cross-cultural investigation of social support and burnout. *European Psychologist* **7**, 256–264.
- Poghosyan L, Aiken LH and Sloane DM** (2009) Factor structure of the Maslach burnout inventory: an analysis of data from large scale cross-sectional surveys of nurses from eight countries. *International Journal of Nursing Studies* **46**, 894–902.
- Rotenstein LS, Torre M, Ramos MA, Rosales RC, Guille C, Sen S and Mata DA** (2018) Prevalence of burnout among physicians A systematic review. *Jama-Journal of the American Medical Association* **320**, 1131–1150.
- Schaufeli WB and Enzmann D** (1998) *The Burnout Companion to Study and Practice: A critical Analysis*. London: Taylor & Francis.
- Schaufeli WB and Van Dierendonck D** (1993) The construct validity of two burnout measures. *Journal of Organizational Behavior* **14**, 631–647.
- Schilling R, Colledge F, Brand S, Ludyga S and Gerber M** (2019) Psychometric properties and convergent validity of the Shirom-Melamed burnout measure in two German-speaking samples of adult workers and police officers. *Frontiers in Psychiatry* **10**, 536.
- Stamm BH** (2010) *The Concise ProQOL Manual, 2nd Ed.* Pocatello, ID: ProQOL.org.
- Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, Bouter LM, De Vet HCW and Mokkink LB** (2018) COSMIN Methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research* **27**, 1159–1170.
- West CP, Dyrbye LN, Satele DV, Sloan JA and Shanafelt TD** (2012) Concurrent validity of single-item measures of emotional exhaustion and depersonalization in burnout assessment. *Journal of General Internal Medicine* **27**, 1445–1452.
- Williamson K, Lank PM, Cheema N, Hartman N and Lovell EO and Emergency Medicine Education Research Alliance (EMERA)** (2018) Comparing the Maslach burnout inventory to other well-being instruments in emergency medicine residents. *Journal of Graduate Medical Education* **10**, 532–536.
- World Health Organisation** (2019) *International Classification of Diseases for Mortality and Morbidity Statistics (11th Revision)*. Retrieved from <https://icd.who.int/browse11/l-m/en>.
- Yadama GN and Drake B** (1995) Confirmatory factor analysis of the Maslach burnout inventory. *Social Work Research* **19**, 184–192.