Valid and Reliable Survey Instruments to Measure Burnout

A key organizational strategy to improving clinician well-being is to measure it, develop and implement interventions, and then re-measure it. A variety of dimensions of clinician well-being can be measured including burnout, engagement, and professional satisfaction. Below is a summary of established tools to measure burnout. Each tool has advantages and disadvantages and some are more appropriate for specific populations or settings. This information is being provided by the Research, Data, and Metrics Working Group of the National Academy of Medicine Action Collaborative on Clinician Well-Being and Resilience.

Click or scroll below for an overview of each valid and reliable instrument to measure burnout, well-being, and other work-related dimensions.

**Burnout**

- Maslach Burnout Inventory – Human Services Survey for Medical Personnel
- Oldenburg Inventory
- Physician Work-Life Study’s Single-Item
- Copenhagen Burnout Inventory

**Composite Well-Being**

- Well-Being Index

**Depression and Suicide Risk**

- The Patient Health Questionnaire-9 (PHQ-9)

**Frequently Asked Questions**
**Purpose**
To measure burnout in individuals who work with people (human services professions).

**Format/Data Source**
Maslach Burnout Inventory – Human Services Survey (MBI-HSS) is a 22-item survey that covers 3 areas: Emotional exhaustion (EE), Depersonalization (DP), and Low sense of personal accomplishment (PA). There are multiple questions for each of these subscales and responses are in the form of a frequency rating scale (never, a few times a year or less, once a month or less, a few times a month, once a week, a few times a week, every day).

**Date**
Measure released in 1981.

**Measure Item Mapping**
- Emotional Exhaustion: 1, 2, 3, 6, 8*, 13, 14, 16, 20
- Depersonalization: 5, 10*, 11, 15, 22
- Personal Accomplishment: 4, 7, 9, 12, 17, 18, 19, 21

*items with highest factor loading on the subscale in samples of physicians and trainees (see below development and testing).\(^2,3\)

**Data Analysis**
It is preferred to examine relationships with subscale scores as continuous variables and outcomes. An alternative approach is to categorize results separately using established definitions of low, average, and high cut-off scores for each subscale (see MBI manual).\(^1\) Investigators often dichotomize into burnout/no burnout but there is no accepted standard definition.\(^4\) A common approach is to consider individuals as having at least one symptom of burnout if they have high scores in either the EE (score of 27 or higher) or DP (score of 10 or higher) subscales. Evidence indicates that high scores on these subscales can distinguish the clinically burned out from the non-burned out.\(^5\) Because this approach identifies individuals whose degree of burnout places them at increased risk of potentially serious personal and professional consequences.\(^6-11\) An alternative option is to consider individuals to have burnout if they have a high EE score along with either a high DP score or a low PA score (score less than 33).\(^4\)

**Development and Testing**
The instrument was developed following exploratory research with interview and questionnaire data, testing in a variety of health and service occupations, and factor and confirmatory data analysis. Reliability coefficients, test re-test reliability, convergent validity, and discriminant validity among human service professionals are summarized in the manual.\(^1,12\)

As the 22-item version is not always practical, such as in long surveys assessing multiple different dimensions, investigators have identified 2 single questions with the highest factor loading on the EE and DP subscales in samples of 2248 medical students, 333 internal medicine residents, 465 internal medicine faculty, and 7905 practicing surgeons: item 8 (“I feel burned out from my work”) and item 10 (“I have become more callous toward people since I took this job”).\(^2,3\) These two items correlated strongly with the EE and DP domains as measured by the full MBI in a sample of more than 10,000 trainees and physicians with an area under the receiver operator characteristic curve of 0.94 and 0.93 for EE and DP, respectively, for these single items relative to the full MBI. Concurrent validity for the 2-item instrument has been established by showing, relative to the full MBI, that the single item measures have strong and consistent correlations with key outcomes in medical students, residents, and practicing surgeons (e.g., suicidal ideation, serious thoughts of dropping out of medical school, poor professionalism, major medical errors).\(^2,3\) A response of “a few times a week” or more on single items indicates a high degree of burnout in each domain. In a separate study of 308 rural physicians and advance practice providers, the 2 items (item 8 and item 10) correlated well with their parent subscales of the full MBI (Spearman's \(r = .89\) and \(.81\), \(p < .0001\)). In multivariable models, the EE item predicted high/low EE, whereas the DP item predicted only low DP.\(^13\)

**Links to Outcomes or Health System Characteristics Related to Health Care Professionals**
Substantial data\(^14,15\) supports associations between burnout as measured using the MBI and health care related outcomes (e.g., medical error,\(^6,10,11\) malpractice,\(^16\) suboptimal patient care practices,\(^17\) physician turnover and early retirement,\(^18,19\) and lower medical knowledge\(^20\)), suboptimal professionalism,\(^7,21\) and personal outcomes (e.g., alcohol abuse -\(^22-24\) suicidal ideation,\(^6,9\) and motor vehicle incidents\(^25\)). From a health system characteristics perspective, associations have been found between burnout and practice setting, work hours, clerical burden, and specialty.\(^14,26-29\)

**Country of Origin**
United States of America

**Past or Validated Applications**
- **Patient age:** adults
- **Population:** human service/helping professionals (e.g., teachers, social workers, police officers), including physicians, residents/fellows, medical students, and nurses
  - National benchmark data available:
To measure burnout in any occupational group.

**Purpose**

To measure burnout in any occupational group.

**Notes**

- Multiple language translations are available
- MBI-General Survey (MBI-GS) is applicable to more general, non-social jobs.

**References**

21. Dyrbye LN, West CP, Satele D, Boone S, Sloan J, Shanafelt TD. A national study of medical students’ attitudes toward self-prescribing and responsibility to report impaired colleagues. Acad Med 2015;Published ahead of print on line
25. West CP, Tan AD, Shanafelt TD. Association of resident fatigue and distress with occupational blood and body fluid exposures and motor vehicle incidents. Mayo Clinic Proceedings 2012;87:1138-44.
Oldenburg Burnout Inventory is a 16-item survey with positively and negatively framed items that covers 2 areas: exhaustion (physical, cognitive, and affective aspects) and disengagement from work (negative attitudes toward work objects, work content, or work in general). There are multiple questions for each of these subscales and responses are in the form of a 4 point Likert scale from strongly agree (1) to strongly disagree (4).

Measure

- Exhaustion: 2, 4, 5, 8, 10, 12, 14, 16
- Disengagement: 1, 3, 6, 7, 9, 11, 13, 15

Data Analysis

Each burnout dimension is treated separately as a continuous variable.

Development and Testing

Developed in response to the MBI not having negatively worded items, and based on job demands-resources model where job demands are primarily related to exhaustion and job resources are primarily related to disengagement. Two factor structure has been confirmed in a sample of Dutch workers, and Dutch physicians and US workers whereas a four factor model (exhaustion, energy, disengagement, and engagement) was supported in study of Chinese nurses. There is some evidence of convergent validity of OLBI with a shortened (16-item) version of the MBI-GS in a sample of 2431 US workers and in a sample of Chinese nurses though convergent validity data suggests positively worded items should be dropped. In a study of 232 Greek employees bivariate correlation between OLBI-exhaustion and MBI-GS-emotional exhaustion was 0.6, and the bivariate correlation between OBLI-disengagement and MBI-GS depersonalization was 0.6. In a study of 528 South African employees working in construction, bivariate correlation between OLBI-exhaustion and MBI-GS-emotional exhaustion was 0.6, and the bivariate correlation between OBLI-disengagement and MBI-GS depersonalization was 0.37.

Links to Outcomes or Health System Characteristics Related to Health Care Professionals

Existing data is limited as a majority of studies have included small samples of physicians and other health care providers, and have mostly been conducted outside of the United States. Studies in Swedish nurses and other Swedish public health professionals suggest that OLBI scores predict intent of turnover and lower self-reported mastery of occupational skills. Correlations have also been reported between OLBI scores and self-rated health (n=342 Swedish medical students and n=290 medical residents). In a longitudinal sample of 186 Swedish medical students, end of medical school OLBI-exhaustion and worries about their future endurance/competence predicted 6-10 month postgraduate OLBI-exhaustion.

Country of Origin

Germany

Past or Validated Applications

- **Patient age:** adults
- **Population:** any occupational group
  - National benchmark data not available for US physicians, medical students, or general population.
- **Setting:** any

Cost

$0. Instrument publicly available in appendix of article.

Notes

- Multiple language translations are available

References

6. Demerouti E, Mostert K, Bakker AB. Burnout and work engagement: a thorough investigation of the independency of both constructs. Journal of Occupational...
Purpose
To measure burnout in any occupational group.

Format/Data Source
Single-item. Stem and response items vary in publications. The following item was utilized in Dolan et al.: "Overall, based on your definition of burnout, how would you rate your level of burnout?" Responses, options are (1) "I enjoy my work, I have no symptoms of burnout," (2) "Occasionally I am under stress and I don't always have as much energy as I once did, but I don't feel burned out," (3) "I am definitely burning out and have one or more symptoms of burnout, such as physical and emotional exhaustion," (4) "The symptoms of burnout that I am experiencing won't go away. I think about frustration at work a lot," and (5) "I feel completely burned out and often wonder if I can go on. I am at a point where I may need some changes or may need to seek some sort of help."

Date
Measure released in 2000.

Measure Item Mapping
N/A

Data Analysis
Often dichotomized as no symptoms of burnout (score of 2 or less) vs. 1 or more symptoms (score of 3 or more). These cut-off scores were not established based on validity evidence.

Development and Testing
In a sample of 5400, VA employees correlation between responses to the PWLS single-item with single-item for MBI-EE (item 8: "I feel burned out from my work") score was r = 0.79. Compared to single MBI-EE item, the PWLS single-item had a sensitivity of 83.2%, specificity of 87.4%, and AUC was 0.93. In a separate sample of 307 physicians PWLS single-item correlated modestly with MBI-EE score r = 0.64. In a third study, PWLS single-item responses in sample of 308 rural physicians and advance practice providers correlated with full MBI EE and DP domain scores (Spearman's r = .72 and .41, p<.0001). In multivariable models, PWLS single item predicted high EE (but neither low EE nor low/high DP) as measured by the MBI. In this sample, the original MBI 2 items (item 8: "I feel burned out from my work" and item 10: "I have become more callous toward people since I took this job") correlated better with their respective parent subscale (Spearman's r = .89 and .81, p<.0001). The summary from that study was that the PWLS single item predicts high levels of EE but not low EE or DP, and that it is not effective at capturing individuals who have evidence of burnout in the depersonalization or personal accomplishment domains.

Links to Outcomes or Health System Characteristics Related to Health Care Professionals
In a study of 422 primary care physicians, PWLS single item burnout characterization was associated with lower satisfaction, greater time pressure, poor work control, and intent to leave the medical practice on univariate analysis. No relationship was found between burnout and quality of care, as measured by chart review of 1419 patients. In a related study involving 426 primary care physicians structural equation modeling found significant and small to modest path coefficients between stress, satisfaction, and PWLS single item burnout and between PWLS single item burnout and self-reported medical error and suboptimal patient care practices.

Country of Origin
United States of America

Past or Validated Applications
- Patient age: adults
- Population: Physicians
- National benchmark data not available for US physicians, medical students, or general population. Some data in VA primary care, including 1769
- **Setting:** any health care setting

**Cost**

$0. Publicly available. 

References


**Measure Item Mapping**

- Overall physical and psychological fatigue: 6 items
- Physical and psychological fatigue related to work: 7 items
- Client-related burnout: 6 items

*(Questions are to be mixed with questions on other topics to avoid stereotyped response patterns)*

**Data Analysis**

Each dimension is separately treated as a continuous variable. The response options are recoded into scores of 100, 75, 50, 25, and 0. Next, items within the subscale are averaged, with one item reverse scored. Higher scores indicate a higher degree of burnout. Possible score ranges for all scales is 0-100. In one study investigators chose a score of 50 or higher to indicate burnout as a dichotomous variable. In a separate study investigators chose scores of 25 or lower, 25 to 50, and higher than 50 to categorize low, intermediate, and high burnout. These cut-off scores were not established based on validity evidence.

**Development and Testing**

Developed with a framework that characterizes the core of burnout as fatigue and exhaustion, which are attributed to specific domains in a person's life (personal, work-related, and client-related). In a sample of 1914 individuals from seven different workplaces CBI scales had high internal reliability, scores correlated with SF-36 scales, and scores predicted future sickness absence, intention to quit, and sleep problems.
Existing data is limited as a majority of studies have included small samples of physicians and other health care providers, and have mostly been conducted abroad. In terms of potential health care related outcomes, CBI scores have been associated with lower perceptions of quality of care (psychosocial care, diagnosis/therapy, quality assurance, diagnostic and therapeutic errors in a study of 1311 German surgeons), nurse turnover intention (in a study of 159 ICU nurses in Iran), self-reported sick absences (prospective study of 824 Danish workers in human service sectors), and sickness days, sleep problems, use of pain killers, and intention to quit work (prospective study of 1914 Danish employees in human sector). In terms of personal outcomes, CBI scores predicted the WHO-Five Well-Being Index score among 317 Canadian residents, and antidepressant treatment, especially among men (prospective study of 2936 Danish employees). From a health system characteristics perspective, associations have been found between CBI score and job strain, over-commitment, and low social support (Taiwanese health care professionals) and between practice setting and recent reorganization at work (598 Norwegian midwives).

**Country of Origin**

Denmark

**Past or Validated Applications**

- **Patient age:** adults
- **Population:** any occupational group
  - National benchmark data not available for US physicians, medical students, or general population.
- **Setting:** any

**Cost**

$0. Publicly available in Table S1 and [http://www.arbejdsmiljoforskning.dk/upload/cbi-scales.pdf](http://www.arbejdsmiljoforskning.dk/upload/cbi-scales.pdf)

**Notes**

- Multiple language translations are available

**References**


**Purpose**

To identify distress in a variety of dimensions (burnout, fatigue, low mental/physical quality of life, depression, anxiety/stress).

**Format/Data Source**

7 or 9-item instrument with yes/no response categories.

**Date**


Data Analysis

A total score is calculated by adding the number of 'yes' responses. In a sample of physicians, medical students, and US workers, every one point increase in score resulted in a step-wise increased probability of distress and risk for adverse personal and professional consequence. For the 7-item version, score range is 0 to 7, and threshold score to identify individuals in distress is 4 or higher for medical students, 5 or higher for residents, 4 or higher for practicing physicians, and 2 or higher for other US workers. In the expanded 9-item version, the original 7-items are scored in a traditional manner, with responses to meaning in work and satisfaction with work-life balance items resulting in 1 point being added or subtracted, resulting in a score range of -2 to 9.

Development and Testing

The 7-item WBI was originally designed to be used in medical students. Development involved input from experts, correlation analysis from previously administered assessments, and a multi-step validation process. After initial development in a sample of 2230 medical students, the efficacy of the WBI was confirmed in a separate sample of 2682 medical students. At a threshold score of 4 or higher, the WBI's specificity for identifying medical students with severe distress ranged from 88-91% with sensitivity of 59-93%. The WBI was validated in a national sample of 7560 US residents in 2012. At a threshold score of 5 or higher the index's specificity for identifying residents with low mental QOL, high fatigue, or recent suicidal ideation was 84%. The score also stratified residents' self-reported medical errors. The WBI was also validated in a national sample of 6994 US physicians. At a threshold score of 4 or higher, the index's specificity for identifying physicians with low mental QOL, high fatigue, or recent suicidal ideation was 86%. The score also stratified career satisfaction, reported intent to leave the current practice, and self-reported medical errors. In 2014, the 7-item WBI was tested in a sample of 5392 US workers and 6880 US physicians, and the 9-item WBI was developed and tested. The 9-item was created in an effort to identify individuals who were thriving, and included items exploring satisfaction with work-life integration and meaning in work, both of which may mitigate the relationship between job-related stress and psychological distress. The 9-item WBI predicted low and high QOL, high fatigue, recent thoughts of suicidal ideation, and burnout in both samples. The area under the curve of the 7-item and the 9-item for identifying burnout was 0.84 and 0.85 in the physician sample, respectively.

Links to Outcomes or Health System Characteristics Related to Health Care Professionals

National studies have found associations between WBI scores and health care related outcomes (e.g., medical error, physician turnover) and personal outcomes (e.g., fatigue, recent suicidal ideation).

Country of Origin

USA

Past or Validated Applications

- **Patient age:** adults
- **Population:** any occupational group
  - National benchmark data available for US physicians, residents, medical students, and general population, with national benchmarks soon available for US nurses and advance practice providers.
- **Setting:** any

Cost

Varies. Information regarding cost and permission to use the tool available at https://www.mededwebs.com/well-being-index.

References

**Purpose**

To measure major depression and suicidal ideation.

**Format/Data Source**

The Patient Health Questionnaire-9 (PHQ-9) is the self-report component of the PRIME-MD (Primary Care Evaluation of Mental Disorders) inventory\(^1\). For each of the 9 DSM-V (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]) depressive symptoms, participants indicate whether, during the previous 2 weeks, the symptom has bothered them “not at all,” for “several days,” for “more than half the days,” or “nearly every day.”

Suicidal ideation is screened for with item 9 of the Patient Health Questionnaire–9 (PHQ-9) (i.e., “Thoughts that you would be better off dead, or hurting yourself in some way” over the past 2 weeks). Positive response to this item increases the cumulative risk for a suicide attempt and suicide completion over the next year by 10- and 100-fold, respectively\(^2\).

**Date**

Measure released in 1999.

**Measures Item Mapping**

One item each for:

1. Interest
2. Mood
3. Sleep
4. Energy
5. Appetite
6. Self-worth
7. Concentration
8. Psychomotor slowing or activation
9. Suicidal ideation

**Data Analysis**

The PHQ-9 is most often used as a continuous measure, with scores for individual items summed to produce a composite depressive symptom score between 0-27. Cut points of 5, 10, 15 and 20 representing mild, moderate, moderately severe and severe levels of depressive symptoms. The PHQ-9 can also be used as a diagnostic algorithm to make a probable diagnosis of major depressive disorder (MDD)\(^3\).

**Development and Testing**

PHQ-9 scores ≥10 have a sensitivity and specificity of 88% for major depressive disorder\(^4\). The PHQ-9 performs similarly across sex\(^5\), age\(^6\) and racial/ethnic groups\(^7\). Importantly for longitudinal assessments, the PHQ-9 shows high sensitivity to change over time\(^5,10\).

Compared to other available depression measures, the PHQ-9 is relatively short and demonstrates good validity, sensitivity and specificity in both clinical and non-clinical populations\(^11\). Further, the PHQ-9 is the primary depression instrument utilized by large health care providers such as the Veterans Administration and the National Health Services, and is the instrument that web users are taken to after a Google search for “clinical depression.”\(^12,13\) (https://www.blog.google/products/search/learning-more-about-clinical-depression-phq-9-questionnaire/). The widespread use of the PHQ-9 ensures a range of normative data for comparison.

**Links to Outcomes or Health System Characteristics Related to Health Care Professionals**

In physicians, PHQ-9 scores have been associated with medical errors, work hours and productivity\(^10,14,15\).

**Country of Origin**

United States of America

**Past or Validated Applications**

- **Patient age:** Adolescents, adults, and older adults
- **Population:** any occupational group
  - From meta-analyses, comparison data are available for the general population, medical students (N=10,386),\(^16\) and resident physicians (N=3,756)\(^17\)
- **Setting:** any

**Cost**
Alternate Depression Measure

The abbreviated 2-item PHQ-2 instrument has been developed for situations where administration of the full PHQ-9 is not feasible. The PHQ-2 is composed of the first two items of the PHQ-9 (assessing low mood and loss of interest) and subjects receive a score between 0 and 3 on each item. With a composite score range between 0-6, scores of ≤2 or ≥3 have been considered a positive screen for depression depending on the study. A positive PHQ-2 screen for depression correlates well with positive screens on the PHQ-9 and other longer depression instruments. Further, the PHQ-2 has generally shown moderate to good sensitivity to detect clinical depression. However, the specificity of PHQ-2 has been variable across studies and low in many studies. Thus, the PHQ-2 is most accurately viewed as a screening tool for depression rather than a diagnostic instrument.

References

22. Wilson R, Agius M. Is there any advantage to using two of the PHQ-2? Are there clear advantages for using one instrument versus another? For example, does use of MBI better enable comparisons with results from previous studies of US physicians or in other specialties?

Notes

- Multiple language translations are available.

The Maslach Burnout Inventory is the gold standard for research purposes. Use of the full MBI allows for scores to be compared to results from previous studies of US physicians (Mayo Clinic Proc, December 2015;90(12):1600-1613). The full MBI is 22-items long and therefore may not be practical in all settings. The use of 2 single items from the Maslach Burnout inventory is the second best option: item 8 ("I feel burned out from my work") and item 10 ("I have become more callous toward people since I took this job") correlate strongly with the emotional exhaustion and depersonalization subscale scores and concurrent validity has also been demonstrated (J Gen Intern Med 2012;27:1445-52. J Gen Intern Med 2009;24:1318-21).
Organizations interested in measuring physician well-being could consider a variety of dimensions including: burnout, stress, fatigue, satisfaction, and quality of life. Various instruments are available to measure these domains. Instruments with national benchmark data and shown to correlate with patient satisfaction, safety, quality measures, productivity, turnover, and other outcomes of interest are preferred. A review of self-reported measures for assessing well-being has been published in BMJ (Linton M, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ Open 2016;6:e010641. doi: 10.1136/bmjopen-2015-010641). Commonly used instruments, number of items, whether or not there are national benchmarks for US physicians and scores have been shown to correlate with relevant outcomes can also be found in the Table on page 6 of the article “Executive Leadership and Physician Well-Being: Nine Organizational Strategies to Promote Engagement and Reduce Burnout” By Drs. Shanafelt and Noseworthy (Mayo Clin Proc. 2017 Jan;92(1):129-146. doi: 10.1016/j.mayocp.2016.10.004).

Conducting a survey to measure burnout or other dimensions of distress is the first step to managing the problem. Including questions on such surveys that explore key drivers of burnout, such as meaning in work, work-load, work efficiency, social support at work, control/flexibility, work-life balance, and organizational culture and values, can provide a starting point for conversation and action.

Are there best practices for combining different instruments?

Surveys are an important research methodology. The most important piece is to use instruments with acceptable reliability and validity that have national benchmark data to help with interpreting results, and are preferably associated with outcomes of interests. Using the entire instrument with exactly the same instructions and response categories is critical. If there are plans to repeat the survey over time the instruments should follow one another in the same order. It is acceptable to include more than one instrument in a survey. Just be sure to keep the instructions that precede the items and the response categories the same as the original instrument.

Survey-related books:


Is there an optimal number or upper limit on the number of questions, to maximize response rates?

There are many factors that influence response rate, including interest of your sample population in the topic and survey length. Short, simple questionnaires typically have better response rates than long, complex surveys. There are a number of factors beyond how many questions there are on a survey that influence how long it takes to complete a survey. Surveys that take less than 10 minutes for an individual to complete tend to have better response rates. Prior to administering a survey to a large group of individuals, it is advised to pilot test the instrument with a smaller cohort. Doing so can help identify errors and ensure the success of the project.

For more details see:


Are there unique considerations for inclusion of depression-related questions in surveys?

When assessing sensitive topics such as depression, suicide, and substance abuse there is potential tension between the desire to directly help respondents exhibiting signs of pathology and maintaining respondent confidentiality. Especially, with the unfortunate continued stigma around mental health among clinicians ensuring confidentiality is critical to respondents and to collecting accurate results. One compromise solution is to take all participants to a new screen at the end of the survey that provides a) general information about depression b) encouragement to seek help if experiencing depressive symptoms and STB c) and information on resources for mental health services by state, including a suicide hotline.

Learn more about the Action Collaborative on Clinician Well-Being and Resilience

The Action Collaborative on Clinician Well-Being and Resilience aims to improve baseline understanding of challenges to clinician well-being, to raise the visibility of clinician stress and burnout, and to elevate evidence-based multidisciplinary solutions. To learn more about their work, please click here.