

3 Good practices for measuring population mental health in household surveys

All OECD countries currently measure population mental health, yet use a variety of tools to capture a multitude of outcomes. In order to improve harmonisation, this chapter poses a series of questions that highlight the criteria to be considered when choosing appropriate survey tools. These criteria include statistical quality, practicalities of fieldwork and data analysis. Overall, there is strong evidence supporting the statistical properties of the most commonly used screening tools for the composite scales of mental ill-health and positive mental health. Four concrete tools (the PHQ-4, the WHO-5 or SWEMWBS, and a question on general mental health status) that capture outcomes across the mental health spectrum are suggested for inclusion in household surveys in addition to already ongoing data collection efforts.

Countries across the OECD are already implementing a variety of survey tools to measure aspects of population mental health. Chapter 2 highlighted that while there is some degree of harmonisation for outcomes such as risk for depression, life evaluation and general psychological distress, there are gaps in coverage for others: in particular, anxiety; other specific mental disorders (bipolar disorder, PTSD, eating disorders and so on); and affect and eudaimonic aspects of positive mental health. Before settling on a concrete list of recommendations for member countries, this chapter provides an overview of the properties that should be considered when selecting a specific tool to measure these outcomes in household surveys.

While OECD countries are already using a variety of tools, including structured interviews, data on previous diagnoses, experienced symptoms and questions on suicidal ideation and suicide attempts, this chapter will focus on the statistical qualities of three other tools in common use: screening tools¹, positive mental health indicators and general questions on mental health status. This is for two reasons. First, these tools are standardised in terms of question formulation and thus provide the easiest foundation on which to make harmonised recommendations. Second, these tools are more commonly featured in general social surveys (as compared to tools for diagnoses or experienced symptoms), which tend to be collected more frequently than health-specific surveys. Taken together, these three tools also provide a holistic measure of mental health, encompassing the full possibility of outcomes conceptualised by both the single and dual continua (see Figure 1.2), and they provide more nuance than, say, measures of suicidal ideation or attempts. These tools are then the most promising when thinking of pragmatic recommendations that can be taken up by the largest number of countries.

When selecting an appropriate tool, the overarching consideration is how to measure the different facets of mental health most accurately – across countries, groups and time – in a way that can be used by government as a part of an integrated policy approach to mental health. High-quality data are needed to provide insights into how societal conditions (economic, social, environmental) affect the mental health of different population groups and whether these conditions contribute to improving or declining mental health. No data are completely without measurement bias, and it is always important that data collection entities enact rigorous quality controls to minimise the amount of noise in a measure. However, there are challenges specific to the measurement of mental health, due to stigma and bias affecting survey response behaviour, different cultural views and evolving attitudes towards mental health over time. Furthermore, household surveys by definition exclude institutionalised populations, including those in long-term care facilities, hospitals or prisons, as well as people with no permanent addresses, all of whom may have higher-than-average risk for some mental health conditions.

Good practices for measuring mental health at the population level differ in several ways from those for measuring mental health at the clinical level. For national statistical offices or health ministries conducting large-scale, nationally representative surveys, implementing long structured interviews is impractical, even though these may be considered the gold standard from a clinical perspective. The end users of the data are different, and policy makers have other needs than clinicians: tracking overall trends (over time, across at-risk groups, among countries), and factors of risk and resilience in population groups vs. diagnosing an individual and developing a treatment plan. These needs guide this chapter's discussion.

This chapter provides a guide to good practices in producing high-quality data on population mental health outcomes, by posing a series of questions for data collectors to consider. High-level findings from this exercise are shown in Table 3.1, below. The specific screening and composite-scale tools included in the table are those that are used most frequently across OECD countries (for more information on each, refer to Table 2.7, Table 2.11, Table 2.12 and Annex 2.B).² Questions are grouped into three overarching categories, covering (1) statistical quality, (2) data collection procedures and (3) analysis. Evidence from existing research is used to illustrate each question area, rather than to comprehensively assess every mental health tool used by OECD countries. These framing questions serve as a lens for assessing the advantages and disadvantages of different tools for measuring population mental health and to guide the concrete recommendations for tool take-up and harmonisation outlined in the conclusion.

Table 3.1. Overview of mental health tool performance on statistical quality, data collection and analysis metrics

Tool Information			Statistical Quality					Data Analysis		Country Coverage
Name	Topic coverage and item length	Reference period	Reliability	Validity	Low missing rates	High comparability across groups	Sensitive to change	Normal distribution	Sensitivity/specificity of thresholds	OECD countries reporting its use
Validated screening tools for assessing mental ill-health										
<i>Psychological distress</i>										
General Health Questionnaire (GHQ-12)	Negative and positive affect, somatic symptoms, functional impairment; 12 items	Recently	✓	×	×	~		×	×	5 of 37
Kessler Scale 6 (K6)	Negative affect; 6 items	Past 4 weeks	○	✓		~			✓	4 of 37
Kessler Scale 10 (K10)	Negative affect, functional impairment; 10 items	Past 4 weeks	✓	✓		~			✓	4 of 37
Mental Health Inventory 5 (MHI-5)	Negative and positive affect; 5 items	Past month	✓	✓				×	~	28 of 37
<i>Depressive symptoms</i>										
Patient Health Questionnaire -8 or -9 (PHQ-8 / PHQ-9)	Negative affect, anhedonia, somatic symptoms, functional impairment (matched to major depressive disorder per DSM-IV and DSM-5 criteria); 8 or 9 items	Past 2 weeks	✓	✓		✓	✓		✓	30 of 37
Patient Health Questionnaire -2 (PHQ-2)	Negative affect, anhedonia; 2 items	Past 2 weeks	~	~		~	✓			8 of 37
Center for Epidemiological Studies Depression Scale (CES-D)	Negative affect, anhedonia; 20 items	Past week	✓	~		×			×	1 of 37
<i>Symptoms of anxiety</i>										
Generalised Anxiety Disorder-7 (GAD-7)	Negative affect, somatic symptoms, functional impairment; 7 items	Past 2 weeks	✓	✓		~	○		~	11 of 37

Generalised Anxiety Disorder-2 (GAD-2)	Negative affect, functional impairment; 2 items	Past 2 weeks	✓	✓		~			~	7 of 37
<i>Symptoms of depression and anxiety</i>										
Patient Health Questionnaire -4 (PHQ-4)	Negative affect, anhedonia, functional impairment; 4 items	Past 2 weeks	~	~		~	~		~	13 of 37
Standardised tools for assessing positive mental health										
Short Form Health Status (SF-12)	Negative and positive affect, functional impairment (Mental Health Component Summary); 12 items	Past 4 weeks	✓	✓		×				8* of 37
Warwick-Edinburgh Mental Well-Being Scale (WEMWBS)	Positive affect, eudaimonia, social well-being; 14 items	Past 2 weeks	✓	✓	✓	~	✓	✓	~	2 of 37
Short Warwick-Edinburgh Mental Well-Being Scale (SWEMWBS)	Positive affect, eudaimonia, social well-being; 7 items	Past 2 weeks	✓	✓	✓	✓	✓	✓	~	6 of 37
WHO-5 Wellbeing Index (WHO-5)	Positive affect; 5 items	Past 2 weeks	✓	✓		✓				6 of 37
Mental Health Continuum Short-Form (MHC-SF)	Positive affect, eudaimonia, life satisfaction, social well-being 14 items	Past month	✓	~		~			×	2 of 37
Single-question self-reported general mental health status										
Self-reported mental health (SRMH)	Varies widely, including self-reported: general mental health status; number of mentally healthy days; recovery from mental health condition; satisfaction with mental health; extent to which mental health interferes in daily life; Single question	Varied (ranges from current assessment to last 12 months)	~	~		×	○			23 of 37

Note: ✓ indicates that the evidence shows this tool performs well on this dimension; ~ indicates that the evidence shows this tool performs only fairly; × indicates that the evidence shows this tool performs poorly; and ○ indicates that evidence is limited or missing. If a cell is blank, this means that no research on this tool / topic combination was reviewed for this publication. * Refers to the fact that Germany included the longer SF-36 (rather than the shorter SF-12) in its 1998 German National Health Interview and Examination Survey, however the instrument will not be used in future due to licensing fees. Refer to Annex 2.A and Annex 2.B for more information about each tool. Country coverage refers to all OECD countries except Estonia, which did not participate in the questionnaire.

Source: Literature reviewed in this chapter; Responses to a questionnaire sent to national statistical offices in January 2022.

Statistical quality

A suitable measurement instrument for population mental health should perform well across a range of statistical qualities, including reliability, validity, ability to differentiate between different latent constructs, minimal non-response or refusals, comparability across groups and sensitivity to change. In addition, practical considerations surrounding a tool are important, such as keeping it short enough in length, with low redundancy between question items, so as to avoid respondent fatigue. These qualities interact with one another, meaning that in practice the goal is to balance the trade-offs of each in order to find a sensible solution. An instrument that performs well in one quality criterion – i.e. validity – may perform poorly in another – i.e. length of the questionnaire and/or non-response rates. Thus before choosing a metric, it is important for survey producers to weigh the costs and benefits of each approach to identify a tool suitable for their context.

How reliable are survey measures of mental health?

Measures of population mental health should produce consistent results when an individual is interviewed or assessed under a given set of circumstances. This concept, called reliability, is about ensuring that any changes detected in outcomes have a low likelihood of being due to problems with the tool itself – i.e. measurement error – and instead reflect actual underlying changes in the individual's mental health (Box 3.1).

Box 3.1. Statistical definitions: Reliability

Two important aspects of reliability are test-retest reliability and internal consistency reliability (OECD, 2013^[1]; OECD, 2017^[2]).

Test-retest reliability concerns a scale's stability over time. A respondent is re-interviewed or re-assessed after a period of time has passed, and their responses to a given questionnaire item are compared to one another. The expectation is that (assuming no change in the underlying state being measured) a reliable measure should lead to responses that are highly correlated with one another. There is no fixed rule for the length of time between the initial interview and follow-up: practice ranges from as short as 2-14 days to six months, depending on the assessment type (NHS Health Scotland, 2008^[3]).

The test-retest criterion must be applied thoughtfully in the case of mental health measurement instruments, as mental health states (and particularly affective states) can fluctuate over short periods of time for a given individual. This means that measurement instruments addressing *specific symptoms* or *states* can be highly reliable yet still produce different results for the same individual over a period of days or weeks, as symptoms and experiences themselves wax and wane. In the context of measuring *population* mental health outcomes, then, test-retest reliability is particularly relevant for:

- Simple measures that concern whether an individual has ever been diagnosed with a mental health condition (where a good instrument should have a very high test-retest correlation)
- Establishing whether a short-form measure (or a measure being validated against a clinical diagnosis) is performing with the same test-retest accuracy as a long-form measure (or clinical diagnosis) when the two are administered to the same respondent, and/or
- Establishing the broad stability of symptom-based measurement scales over short time periods and across large samples - i.e. while the test-retest correlation of questions for a set of symptoms is unlikely to be perfect for a given individual (if symptoms themselves are not always

stable), day-to-day fluctuations in symptoms at the individual level can be expected to wash out across large samples to produce a similar distribution of scores over a short time period.²

Assessing test-retest reliability therefore indicates a trade-off between measures that are sufficiently stable, yet sensitive to change over time. An instrument that performs well on test-retest reliability may perform poorly on tests to measure sensitivity to change, which underscores the importance of looking at statistical quality measures holistically when making decisions as to which tools to implement.

Internal consistency reliability assesses the extent to which individual items within a survey tool are correlated to one another when those items aim to capture the same target construct. In the context of measuring population mental health, this might mean that, in a battery of items designed to measure depression and anxiety, the depression items correlate with one another, and the anxiety items correlate with one another (see also Box 3.3 for a discussion of factorial validity). The most widely used coefficient for internal consistency reliability is Cronbach's alpha, which is a function of the total number of question items, the covariance between pairs of individual items and the variance of the overall score.¹ Although there is not universal consensus, most researchers agree that a coefficient value between 0.7 and 0.9 is ideal (NHS Health Scotland, 2008^[3]). Values below 0.7 may reflect the fact that items within the scale are not capturing the same underlying phenomenon (OECD, 2013^[1]), while values above 0.9 may indicate that the scale has redundant items.

Notes:

1. The Cronbach coefficient alpha is commonly used in the literature to assess the internal consistency reliability of multi-item tools. The coefficient is calculated by multiplying the mean paired item covariance by the total number of items included in the scale and dividing this result by the sum of all elements in the variance-covariance matrix (OECD, 2013^[1]). This results in a coefficient ranging from 0 (scale items are completely independent from one another, no covariance) to 1 (scale items overlap, complete covariance).

2. The definition of "a short time period" is subjective and can vary depending on circumstance. For example, although the period of a couple of days may be deemed an acceptably short period of time over which a test-retest assessment could be administered, if there were to be an extreme shock in the intervening days, either positive or negative, there would be good grounds to expect change in the underlying distribution. Frequent data collection on mental health during the COVID-19 pandemic illustrated the volatile nature of many affect-based measures, with large spikes coinciding with the introduction / easing of confinement policies.

The performance of screening tools on measures of reliability varies across tools and the outcomes they measure. There are mixed findings for general measures of psychological distress. The General Health Questionnaire (GHQ-12) as well as the Short Form-36 (SF-36) and its shorter sub-component, the Mental Health Inventory (MHI-5), have been shown to have good reliability (Schmitz, Kruse and Tress, 2001^[4]; Ohno et al., 2017^[5]; Elovainio et al., 2020^[6]; Strand et al., 2003^[7]); however, while the longer Kessler (K10) has been shown to be internally consistent, the test-retest reliability of the shorter Kessler (K6) tool has not been assessed in any studies (El-Den et al., 2018^[8]; Easton et al., 2017^[9]).

Conversely, screening tools for specific mental conditions – especially depression – are the most studied, and they have been shown to be reliable in terms of both test-retest reliability and internal consistency reliability. A meta-analysis of 55 different screening tools for depression found the Patient Health Questionnaire (PHQ-9) to be the most evaluated tool, with a number of studies concluding that both it and the PHQ-8 (a shorter version with the final question on suicidal ideation removed) have high reliability and validity (El-Den et al., 2018^[8]). The same report, however, found that the shorter Patient Health Questionnaire-2 (PHQ-2) lacked consistent data on validity and reliability: among the six reports that evaluated the PHQ-2, only one reported on its internal consistency or test-retest reliability (El-Den et al., 2018^[8]), which led the authors to caution that the reliability of the PHQ-2 cannot be confirmed with available data. The Center for Epidemiological Studies Depression Scale (CES-D), although less studied than the PHQ, has also been found to have good reliability, on both metrics (Ohno et al., 2017^[5]). Among anxiety tools, the Generalised Anxiety Disorder screeners (both the longer GAD-7 and shorter GAD-2) have been found to be reliable, with good test-retest and internal consistency reliability (Ahn, Kim and Choi, 2019^[10]; Spitzer et al., 2006^[11]).

A study of the Patient Health Questionnaire-4 (PHQ-4), which combines the PHQ-2 and GAD-2 to generate a composite measure of both depression and anxiety, found lower, yet still acceptable reliability (Cronbach's alpha > 0.80 for both sub-scales) (Kroenke et al., 2009^[12]). Another study of the PHQ-4 found lower item-intercorrelations but deemed the reliability to be acceptable given the short length of the scales (Löwe et al., 2010^[13]).³ Because Cronbach's alpha is in part a function of the total item length (refer to Box 3.1), shorter scales will perform worse on tests of internal consistency by construction. However shorter measures, with less redundancy between question items, are often preferred by survey creators, as they entail a lower burden for respondents.

Composite scales capturing aspects of positive mental health have also been found to be reliable. A study of the 14-question Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) tool found it to have high test-retest reliability (0.83 at one week) and a high Cronbach's alpha (around 0.9) (Tennant et al., 2007^[14]; NHS Health Scotland, 2016^[15]). The authors cautioned, though, that the high Cronbach's alpha suggests some redundancy in the scale items, a concern that led to the development of the shorter seven-item version (SWEMWBS) (Tennant et al., 2007^[14]; NHS Health Scotland, 2016^[15]). Multiple studies of WEMWBS and SWEMWBS found them both to have strong test-retest reliability (Stewart-Brown, 2021^[16]; Shah et al., 2021^[17]). The World Health Organization-5 (WHO-5) composite scale has also been tested for reliability in a variety of settings (Dadfar et al., 2018^[18]; Garland et al., 2018^[19]). Similarly, the MHC-SF has been found to have high internal reliability, though its test-retest reliability is only moderate (Lamers et al., 2011^[20]).

Fewer studies have investigated the reliability of general self-reported indicators of mental health status; however, evidence from the United States suggests that these measures have acceptable test-retest reliability. The health-related quality-of-life tool used by the United States Centers for Disease Control, the Behavioral Risk Factor Surveillance System (BRFSS) survey, measures perceived health by combining physical and mental health. A study in the state of Missouri found that the shorter version of the tool, with four items, has acceptable test-retest reliability and strong internal validity, although reliability was lower among older adults (Moriarty, Zack and Kobau, 2003^[21]).

Box 3.2. Key messages: Reliability

- Most mental health screening tools, including both surveys that identify specific mental disorders and those that identify positive mental health, have been found to have strong reliability, as measured through both test-retest and internal consistency measures.
- Test-retest reliability must be considered in tandem with a measure's sensitivity to change over time, rather than blindly applied as a quality criterion.
- There is strong evidence for the reliability of screening tools (especially those focusing on depression) and, to a somewhat lesser extent, positive mental health composite scales. However, fewer studies have been done to assess the reliability of general self-reported indicators of mental health status; more research is needed in this area.

How well does the tool measure the targeted outcome?

In addition to being reliable, a good measurement instrument must be valid, i.e. the measures provided by the tool should accurately reflect the underlying concept. For indicators that are more objective, validity can be assessed by comparing the self-reported measure against an objective measure of the same construct. For example, respondents' self-reported earnings could in theory be cross-checked with their tax returns, or pay slips, to ascertain whether their response was reported accurately. Of course there are practical reasons that prevent this from being done systematically, but this illustrates that there are ways of assessing the validity of self-reported earnings data. Conversely, it is not possible to ascertain the "objective truth" of a subjective indicator, such as subjective well-being, trust or indeed mental health. This

does not mean that validity cannot be assessed: OECD measurement guidelines use the concepts of face validity, convergent validity and construct validity to assess the validity of subjective indicators (OECD, 2013^[1]; OECD, 2017^[2]) (Box 3.3).

Unlike many subjective indicators, the bulk of screening tools to assess mental health have been validated against diagnostic interviews for common mental disorders, which provide a rigorous assessment of their accuracy and real-world meaning. The most common diagnostic interview against which mental health screening tools are validated is the World Health Organization's Composite International Diagnostic Interview (WHO-CIDI), which was designed for use in epidemiological studies as well as for clinical and research purposes (see Chapter 2 for more details). This tool allows to measure the prevalence of mental disorders, the severity of these disorders, their impact on home management, work-life balance, relationships and social life, as well as mental health service and medications use. Although the CIDI is widely accepted as a gold standard against which mental health survey items should be assessed, it is not immune to criticisms and validity concerns (Box 3.4).

Box 3.3. Statistical Definitions: Validity

Validity is more difficult to ascertain than reliability, especially for subjective data for which an objective truth is unknowable, and which typically cannot be compared to an equivalent objective measure. Three ways of assessing validity for subjective measures include face validity, convergent validity and construct validity.

Face validity evaluates whether the indicator makes intuitive sense to the respondent and to (potential) data users. One way to indirectly measure face validity is through non-responses. High levels of non-response may indicate that respondents do not understand or see the relevance or usefulness of the question. In the case of mental health, high levels of non-response may also reflect a degree of discomfort with the topic due to stigma and bias, rather than lack of face validity. (An extended discussion of non-response and mental health measures appears later in this chapter.) Cognitive interviewing can also be used.

Convergent validity is assessed by how well the indicator correlates to other proxies of the same underlying outcome. Using mental health tools as an example, were a researcher to introduce a new tool to assess anxiety, s/he could test its convergent validity by comparing it to pre-existing screening tools for data on anxiety, diagnosis or mental health service use, self-reported assessments of anxiety level, and/or bio-physical markers of stress and anxiety (heart rate, blood pressure, neuroimaging, etc.).

Construct validity is the extent to which the indicator performs in accordance with existing theory or literature. For example, research shows that mental health and physical health are correlated with one another and co-move. Therefore, if a new mental health tool showed little correlation with physical health, or if changes in mental health as measured by this tool did not reflect any changes in physical health, the scale would be suspected of having low construct validity. The growing literature on the social determinants of health can also be leveraged to assess construct validity, in a similar way.

In addition to the three aspects of validity mentioned above, clinical validations of mental health survey items often refer to three additional assessments: criterion validity, factorial validity and cross-group validity.

Criterion validity exists only when there is a gold standard against which an item can be compared. In the case of mental health, this gold standard is typically a structured interview (e.g. the CIDI, refer to Annex 2.B). Criterion validity assesses the psychometric properties of a measure, i.e. how it compares to the gold standard. A measure is said to be sensitive if it can accurately identify a "true positive" (i.e. how often the survey accurately identifies someone at risk of, say, depression); it is specific if it can accurately identify a "true negative" (i.e. it accurately identifies someone as *not* at risk for depression).

In order to establish diagnostic accuracy, sensitivity and specificity are plotted in a receiver operating characteristic (ROC) curve at various thresholds. The area under the curve (AUC) can then be used to assess the diagnostic performance of the screening tool in comparison to the gold standard.¹

Factorial validity assesses whether a multi-item survey tool is measuring one, or several, underlying concepts. In almost all cases, unidimensionality is desired if only a single construct is being assessed; this provides assurance that the mental health tool is measuring, for example, depression, anxiety or latent well-being. However, if a scale is assessing multiple dimensions of mental health, then multidimensionality is desired. For example, factor assessments for the PHQ-4, which measures depression and anxiety, indeed identify two latent factors (Löwe et al., 2010^[13]). Factorial validity is commonly assessed using either confirmatory factor analysis (CFA) or exploratory factor analysis (EFA). In the former, researchers test a hypothesis that the relationship between an observed variable (e.g. respondents' answers to the PHQ-8 tool) and an underlying latent construct (e.g. depression) fits a given model. That is, using CFA, researchers test the hypothesis that an observed dataset has a given number of underlying latent factors. Using EFA, researchers do not impose a theoretical model and instead work backwards to uncover the underlying factor structure (Suhr, 2006^[22]).

Cross-group validity, or cultural validity, refers to the extent to which a measure is applicable across different population groups. There are a range of ways that cross-group validity can bias mental health outcome measures, including through cultural factors affecting the way in which symptoms are expressed, clinical bias (either implicit or explicit), language limitations of the respondent (if the tool is being implemented in a language other than their mother tongue) and differences in response behaviour (e.g. greater likelihood to choose midpoint values on Likert scales rather than extreme values) (Leong, Priscilla Lui and Kalibatseva, 2019^[23]). Cross-group validity is best ensured by validating a survey tool in the requisite population, rather than applying it blindly.

Notes:

1. A receiver operating characteristic curve provides a visualisation of diagnostic ability by plotting the true positive rate against the true negative rate. The curve can be used to determine the optimal cut-off point, which minimises both Type 1 (false positive) and Type II (false negative) errors. ROC analysis is used in determining the threshold cut-off scores, which are discussed later in this chapter. For more information on ROC and its use in clinical psychology, refer to (Pintea and Moldovan, 2009^[24]) and (Streiner and Cairney, 2007^[25]).

To assess the validity of screening tools, researchers typically implement a study in which respondents both answer the self-reported scale and participate in a structured CIDI interview, with their responses to both then compared. A screening tool with high sensitivity and specificity is said to have high criterion validity. Although criterion validity ensures that screening tools are designed to mirror diagnostic outcomes from the CIDI, screening tools by design estimate higher prevalence rates for specific mental disorders (see Box 2.1). Convergent validity is assessed by comparing different screening tools against one another to see whether a new tool for measuring, say, depression, performs similarly to existing measures for depression. This approach is often used when testing shortened versions of screening tools, to see whether the truncated survey performs as well as its longer, more in-depth, predecessor. The majority of screening tools described in this chapter have been validated against diagnostic interviews for common mental disorders and have reported good psychometric properties (high sensitivity and specificity) across age groups, gender and socio-economic status (Gill et al., 2007^[26]; O'Connor and Parslow, 2010^[27]; Huang et al., 2006^[28]) (Box 3.3).

Box 3.4. Validity of structured interviews

One important caveat to using structured interviews to validate screening tools is that it presupposes the structured interviews to be an accurate measure of “true” underlying mental health. This issue is raised in two different contexts: (1) most screening tools used in OECD countries were validated against the fourth version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), published in 1994, which is now outdated, rather than against the DSM-5, published in 2013; and (2) the extent to which the DSM itself provides accurate diagnostic data cross-culturally.

The first concern relating to the validity of structured interviews has to do with the fact that none of the screening tools commonly in use have been validated against the newer DSM version (Statistics Canada, 2021^[29]). Yet, in total, there are 464 differences between the DSM-IV and DSM-5. Broadly speaking, the DSM-5 includes fewer diagnostic categories, as many previously separate disorders share a number of features or symptoms. In addition, greater effort was made to separate an individual’s functioning status from their diagnosis. One area that could have an impact moving forward is the lowering of the diagnostic threshold for generalised anxiety disorder – a move that has been criticised by some psychiatrists for pathologising what had previously been considered quotidian worries (Murphy and Hallahan, 2016^[30]).

In sum, even though there are always changes between DSM updates that include the restructuring of diagnostic categories and the updating of some diagnosis criteria, there is by design a degree of continuity between different DSM versions, and most changes are minor. Regardless, in order to be up to date with most recent clinical practice, instruments like the CIDI would benefit from an update.

On the second point, there are concerns about the applicability of these diagnostic validations to non-US regions and population groups, which at the very least would require validation studies to be conducted in different local contexts. Beyond this, validating mental health screening tools in more geographically diverse clinical settings may be insufficient if the clinical diagnoses underpinning the validation are themselves flawed. Haroz and colleagues investigated the extent of this cross-cultural bias by reviewing 138 qualitative studies of depression reflecting 77 different nationalities and ethnicities (Haroz et al., 2017^[31]). They found that only 7 of the 15 most frequently mentioned features of depression across non-Western populations reflect the DSM-5 diagnosis of Major Depressive Disorder. DSM-specified diagnostic features including “problems with concentration” and “psychomotor agitation or slowing” did not appear frequently, while features including “social isolation or loneliness”, “crying”, “anger” and “general pain” – none of which are included as diagnostic criteria – did. Some features arose more frequently in certain regions: “worry” in South and Southeast Asia, and “thinking too much” in Southeast Asia and sub-Saharan Africa. This implies that the close alignment of the PHQ-9 or the GAD-7 with the DSM criteria could in theory limit detection of the underlying targeted construct (i.e. depression, anxiety) relative to longer or more comprehensive screening tools and/or structured interviews (Ali, Ryan and De Silva, 2016^[32]; Sunderland et al., 2019^[33]).

Although criticisms of DSM criteria do exist, the DSM still remains the most useful tool for enabling cross-country comparative data on mental health outcomes. While improvements could be made, the DSM includes considerations of cultural validity in its drafting, which are updated in each subsequent iteration.

Moving beyond clinical psychology, a few OECD countries have expanded their definition of mental health to encompass a wider range of viewpoints, beyond the traditional ones rooted in a Western perspective. In New Zealand, for example, the Government Inquiry into Mental Health and Addiction includes a Māori perspective of mental health (New Zealand Government, 2018^[34]). In a similar vein,

the Swedish government will solicit feedback from the Sami parliament when drafting its upcoming strategy on mental health (Public Health Agency of Sweden, 2022^[35]).

Across measures of general psychological distress, the Kessler and MHI-5 scales have stronger criterion validity than the GHQ-12. Studies have found that the K10 and K6 scales have strong psychometric properties (encompassing both reliability and validity) and better overall discriminatory power than the GHQ-12 in detecting depressive and anxiety disorders (Furukawa et al., 2003^[36]; Cornelius et al., 2013^[37]). The mental health component of the SF-12 tool is also better able to discriminate between those with and those without specific mental health conditions, as compared to the GHQ-12 (Gill et al., 2007^[26]). While the MHI-5 tool has been found to be just as valid as the longer MHI-18 and GHQ-30 to assess a number of mental health conditions, including major depression and anxiety disorders, it performed less well than the MHI-18 for the full range of affective disorders (Berwick et al., 1991^[38]). While the MHI-5 was designed as a general tool, it has been proven effective to identify a specific risk for depression and/or anxiety (Yamazaki, Fukuhara and Green, 2005^[39]; Rivera-Riquelme, Piqueras and Cuijpers, 2019^[40]).

A recent meta-analysis of the sensitivity and specificity of instruments used to diagnose and grade the severity of depression reported that, on average, the PHQ-9 demonstrated the highest sensitivity and specificity relative to other screening tools, including the CES-D (Pettersson et al., 2015^[41]). A different version of the PHQ-8 has been used in the CDC Behavioral Risk Factor Surveillance System (BRFSS) survey. This measure, referred to as the PHQ-8 days, asks respondents how many days over the past four weeks they have experienced each of the eight depressive symptoms that make up the PHQ-8. This yields a scale ranging from 0-112 and can provide a look at depression risk that is more granular – better identifying individuals who may be at risk for mild depression but currently have higher levels of mental well-being – and also more sensitive to change (Dhingra et al., 2011^[42]). The PHQ-2 has been assessed for its internal consistency, construct validity and correlation convergent validity; however, a meta-analysis did not find evidence of studies of criterion validity (El-Den et al., 2018^[8]). Another overview cites evidence for the PHQ-2 as having good criterion validity for specific populations such as older adults, pregnant or post-partum women, and patients with specific conditions such as coronary heart disease or HIV/AIDS (Löwe et al., 2010^[13]).

While self-report scales for depressive symptoms tend to be well validated, scales for anxiety disorders have been found to be somewhat less sensitive and specific in clinical populations. Research suggests this may be because different types of anxiety disorders have more heterogeneous symptoms than depressive disorders (Rose and Devine, 2014^[43]). Despite this, both the GAD-7 and GAD-2 have been validated in a number of studies. The GAD-7 was designed to provide a brief clinical measure of generalised anxiety disorder, and its validation exercise found it to have good validity (criterion, construct, factorial, etc.). Furthermore, factorial validity assessments of the GAD-7 and PHQ-8 found that, despite a high correlation between the anxiety and depression scales (0.75), the two scales are complementary and not duplicative; more than half of patients with high levels of anxiety did not also have high levels of depression (Spitzer et al., 2006^[11]). The high correlations of the GAD-7 with two other anxiety scales indicated good convergent validity (Kroenke et al., 2007^[44]; Spitzer et al., 2006^[11]).⁴ Both the GAD-7 and the shorter GAD-2 perform well in detecting all four major forms of anxiety disorders: generalised anxiety disorder, panic disorder, social anxiety disorder and post-traumatic stress disorder (Kroenke et al., 2007^[44]).

The PHQ-4 has been found to be a valid tool for measuring the combined presence of risks for both depression and anxiety. As noted above, its component parts – the PHQ-2 and GAD-2 – have been validated against diagnostic criterion standard interviews (with caveats to the broader applicability of PHQ-2 criterion validity, as mentioned above). Studies have shown that PHQ-4 scores are associated with the SF-20 functional status scale and health information such as disability days used, etc., providing evidence for convergent and construct validity. Furthermore, factorial analysis has found that the PHQ-4 has a two-

dimensional structure with two discrete factors, picking up on both depression and anxiety disorders (Löwe et al., 2010_[13]).

Composite scales capturing aspects of positive mental health have also been found to have good validity. WEMWBS was found to have good criterion and convergent validity, being highly correlated with other scales that capture positive affect. WEMWBS and the WHO-5 are, unsurprisingly, highly correlated with one another (correlation coefficient of 0.77) (NHS Health Scotland, 2016_[15]), with WEMWBS being slightly less correlated with other measures of mental health that had a stronger focus on physical health or psychological distress (including the GHQ-12). Another study on WEMWBS found that the shorter version of the screening test was highly correlated with the longer version, making it an efficient and quicker alternative to the longer 14-question version (Stewart-Brown et al., 2009_[45]). Despite its length, Rasch analysis has found that WEMWBS is unidimensional with one underlying factor (Stewart-Brown, 2021_[16]).⁵ Multiple studies have shown the MHC-SF has good convergent validity (Guo et al., 2015_[46]; Petrillo et al., 2015_[47]; Lamers et al., 2011_[20]), however cognitive interviews in Denmark found that it had poor face validity, especially for questions on the social subscale (Santini et al., 2020_[48]).

Although designed as measures of positive mental health, both WEMWBS and the WHO-5 have been shown to be effective screeners for depression and/or anxiety. A study found the WHO-5 to have high sensitivity, but low specificity, in identifying patients with depression in a clinical setting (Topp et al., 2015_[49]). A study of SWEMWBS found it to be relatively highly correlated with the PHQ-9 ($\rho = 0.6-0.8$) and the GAD-7 ($\rho = 0.6-0.7$), suggesting that it is an acceptable tool for measuring common mental disorders (CMD); however, other tools may be more sensitive in identifying and distinguishing between individuals with worse levels of mental health (Shah et al., 2021_[17]). A study comparing WEMWBS to the GHQ-12, through multidimensional item-response theory, found that both tools appear to measure the same underlying construct (Böhnke and Croudace, 2016_[50]).

Self-reported mental health (SRMH) indicators have been compared to validated clinical measures of mental health and have been shown to be related to, though distinct from, other mental health scales. SRMH is correlated with the Kessler scales, the PHQ and the mental health component of the SF-12 and is often used in the validation process of other mental health screening tools as a test for convergent validity. Furthermore, SRMH is associated with poor physical health and an increased use of health services. Although related, research has shown that correlations between SRMH and screening tools are moderate, suggesting that they are capturing slightly different phenomena (Ahmad et al., 2014_[51]). The authors note that further research is needed but suggest that findings from longitudinal studies of self-reported *physical* health could shed some light. SRMH measures were shown to be stronger predictors of mortality, morbidity and service use than other indicators, and that SRMH may be capturing mental health problems that do not yet manifest in screening tools (Ahmad et al., 2014_[51]). Conversely, health-related quality of life (HRQoL) – which measures both physical and mental health – has been found to have construct and criterion validity that is good and comparable to the SF-36 scale (Moriarty, Zack and Kobau, 2003_[21]).

Studies of mental health screening tools have yielded conflicting evidence as to whether single-item mental health questions are sufficiently valid. A study assessing the comparative performance of the MHI-5 and MHI-18 (which concluded in favour of the shorter version) found that even a single question – “how often were you feeling downhearted and blue?” – performed as well as the MHI-5, MHI-18 and GHQ-30 at detecting major depression (Berwick et al., 1991_[38]). However, studies assessing ultra-short screening tools found that even two questions perform significantly better at screening for depression than does a single question (Löwe et al., 2010_[13]). Conversely, the Australian Taking the Pulse of the Nation (TPPN) survey, administered throughout the COVID-19 pandemic, found that the psychometric properties of its single-item mental health measure compared favourably to the K6: the items were highly correlated ($\rho = 0.82$), and the single-item measure had high sensitivity for psychological distress (Botha, Butterworth and Wilkins, 2021_[52]).

Box 3.5. Key messages: Validity

- All of the mental health screening tools commonly used by OECD member states have been validated in clinical settings and found to have strong convergent, construct and criterion validity.
- Composite scales for positive mental health have also been found to have strong psychometric properties, and they have proven effective as screeners for specific mental health conditions such as depression and/or anxiety.
- Criterion validity is assessed by the survey tool's performance in comparison to a clinical diagnostic interview gold standard; however, this presupposes the validity of clinical diagnoses, which may not hold in all contexts.

What do non-response rates tell us about stigma? How does this affect the comparability of mental health data across groups?

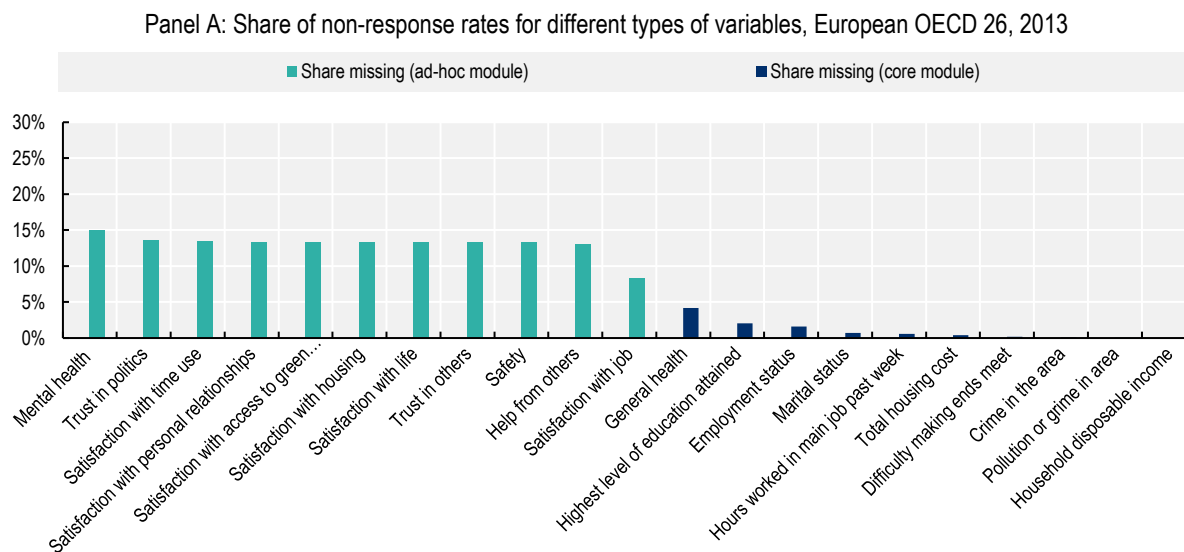
The stigma associated with mental illness may lead to misreporting – and under-reporting – of one's mental health conditions (Hinshaw and Stier, 2008^[53]). Low levels of mental health literacy can also lead to under-reporting, with individuals not recognising their own experienced symptoms as representative of an underlying condition (Tambling, D'Aniello and Russell, 2021^[54]; Dunn et al., 2009^[55]; Coles and Coleman, 2010^[56]).⁶ Feelings of stigma towards mental health conditions remain important in all OECD countries, with large differences between them. A survey conducted in 2019 found that, in 19 OECD countries, 40% of respondents did *not* agree with the statement that mental illness is just like any other illness, and a quarter agreed that anyone with a history of mental disorders should be prevented from running for public office (OECD, 2021^[57]). Because of stigma, respondents may either conceal their true conditions when answering mental health surveys or may choose not to participate in the first place. When administering surveys on sensitive subjects, providing clear assurances of data confidentiality and ensuring that the interview is conducted in a private place, out of hearing of family members, minimise the likelihood of respondent refusal (Singer, Von Thurn and Miller, 1995^[58]; Krumpal, 2013^[59]).⁷

Evidence shows that those experiencing psychological distress or a specific mental disorder are more likely to refuse to participate in a survey; this non-response bias then leads to underestimates of the overall prevalence of mental ill-health (de Graaf et al., 2000^[60]; Eaton et al., 1992^[61]; Mostafa et al., 2021^[62]). A recent study, which compared the effect of psychological distress on a number of economic transitions (e.g. falling into unemployment), using both the GHQ-12 score and a version of it adjusted for misreporting behaviour scores, showed that the original version of the GHQ-12 score underestimated the effect of psychological distress on transitions into better-paid jobs and higher educational attainment (Brown et al., 2018^[63]). Thus, misreporting of symptoms of psychological distress can lead to biased and inconsistent estimates. However, not all studies come to the same conclusion: the US National Comorbidity Survey Replication (NCS-R) study found no evidence of non-response leading to underestimates of disorder prevalence (Kessler et al., 2004^[64]).

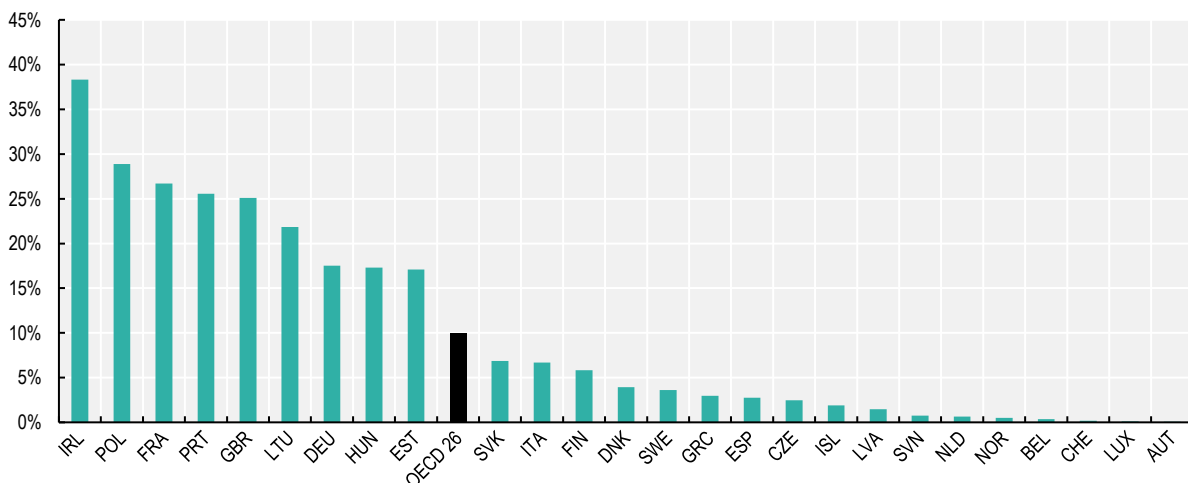
Evidence from the 2013 European Union Statistics on Income and Living Standards (EU-SILC) survey shows that non-response rates for mental health questions are high (15%), but still comparable to those for other subjective variables (e.g. 13% for trust in politics, 8% for satisfaction with one's job (Figure 3.1 Panel A)). High non-response rates for mental health (as measured through the MHI-5) may partly reflect the way in which the EU-SILC survey is implemented. Each year, an ad-hoc module featuring additional questions on a specific topic is implemented in addition to the core module (in 2013 this module focused on well-being), implying that some respondents may have problems in answering questions that were not asked in previous waves.

Missing and non-response rates for mental health variables vary widely between countries (Figure 3.1, Panel B). In the 2013 EU-SILC survey, missing rates for the mental health module were higher than 20% in Ireland, Poland, France, Portugal, the United Kingdom and Lithuania, but were below 1% in Norway, Belgium, Switzerland, Luxembourg and Austria.

Figure 3.1. Non-response rates are higher for mental health questions than they are for other variables, and vary substantially across countries



Panel B: Share of non-response rates for mental health questions (MHI-5), European OECD 26, 2013



Note: This figure only includes individuals who have agreed to participate in the survey, and subsequently choose not to answer individual question items; it does not consider those who refuse to participate in the full survey. A respondent is deemed to be missing mental health data if they refused, or replied “do not know”, to at least four of the five individual items on the MHI-5. Refer to Annex 2.B for more information about specific tools.

Source: OECD calculations based on the *European Union Statistics on Income and Living Conditions (EU-SILC)* (n.d.^[65]) (database), <https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions>.

Table 3.2 shows some suggestive evidence that differences in non-response rates by country could be related to levels of stigma; in nine European OECD countries, the prevalence of any depressive disorder (as measured by the PHQ-8) is inversely correlated with the prevalence of mental health stigma, as measured by the share of the population who agree that people with a history of mental health problems should be excluded from running for office. Therefore, in countries with more stigma, the prevalence of depressive disorder risk is also lower – perhaps because of reluctance to report.⁸

Table 3.2. The relationship between stigma and prevalence is complex, but in some instances, stigma may lead to lower reported prevalence of mental disorders

Correlations between indicators of stigma towards mental health and prevalence of mental health conditions, nine European OECD countries

	Exclude from office if mental health history	Seeking treatment shows strength	Mental health like any other illness	Prevalence of psychological distress (MHI-5)	Prevalence of any depressive disorder (PHQ-8)	Share missing psychological distress responses (MHI-5)
Exclude from office if mental health history	1					
Seeking treatment shows strength	0.11	1				
Mental health like any other illness	0.11	0.56	1			
Prevalence of psychological distress (MHI-5)	0.14	-0.44	-0.47	1		
Prevalence of any depressive disorder (PHQ-8)	-0.83***	-0.19	-0.06	0.13	1	
Share missing psychological distress (MHI-5) responses	0.17	-0.6	0.11	-0.001	0.03	1

Note: Table displays listwise correlations. The three stigma questions ask respondents to indicate the extent to which they agree with the statement. For the first, agreement entails stigma; for the second two, agreement entails the absence of stigma. For details on the MHI-5 and PHQ-8 measures, see Annex 2.B. * Indicates that Pearson's correlation coefficients are significant at the $p < 0.10$ level, ** at the $p < 0.05$ level, and *** at the $p < 0.01$ level. $N = 9$ countries.

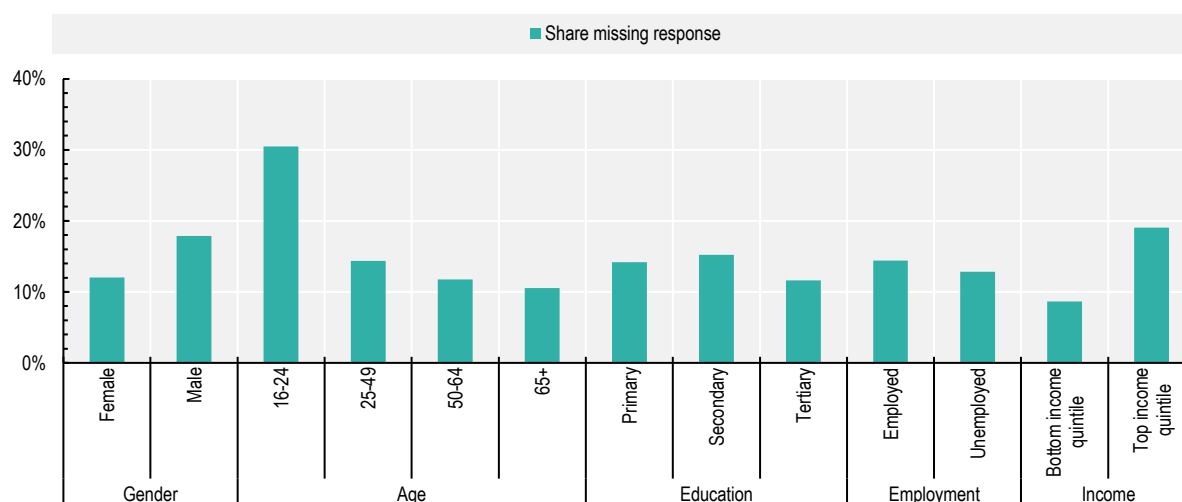
Source: Stigma data originally come from an Ipsos survey, as published in OECD (2021^[57]), *Fitter Minds, Fitter Jobs: From Awareness to Change in Integrated Mental Health, Skills and Work Policies*, Mental Health and Work, OECD Publishing, Paris, <https://dx.doi.org/10.1787/a0815d0f-en>; MHI-5 data come from OECD calculations based on the 2018 *European Union Statistics on Income and Living Conditions* (EU-SILC) (n.d.^[65]) (database), <https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions>; PHQ-8 come from OECD calculations based on *European Health Interview Survey* (EHIS) wave 3 data (n.d.^[66]) (database), [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_\(EHIS\)](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_(EHIS)).

In order to understand what bias is introduced by non-response rates, it is important to understand the profile of those who are choosing not to respond to mental health questions. Figure 3.2 shows these shares for a number of socio-demographic groups – gender, age cohort, education level, labour market status and income quintile. Panel A displays non-responses to mental health questions for 26 European OECD countries, while Panel B shows those for the United Kingdom. For both data sources, women, those with higher levels of education, and older age cohorts are more likely to answer mental health questions, while men, young people and those with lower levels of education are more likely to not respond. These results are in line with a report describing stigma towards mental health in Sweden: women were found to be more likely than men to have positive feelings towards those with mental health conditions, while young people

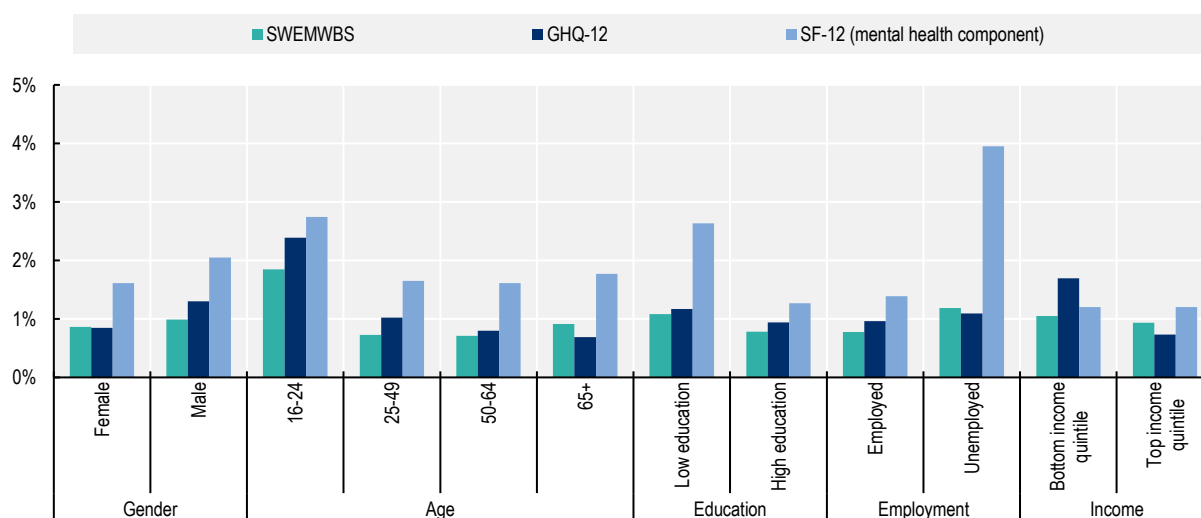
were more likely than older people to report that it would be difficult to talk about their own mental illness with someone else (Folkhälsomyndigheten, 2022^[67]). In European countries, there is a clear difference by income – those in the top income quintile are less likely to respond – however, this pattern does not hold for the United Kingdom. A study on non-response rates in longitudinal health surveys among the elderly in Australia found that those with lower occupational status and less education were less likely to participate (Jacomb et al., 2002^[68]); however, neither risk for depression nor anxiety influenced refusal rates. The Netherlands Mental Health Survey and Incidence Study-2 (MENESIS-2) found higher non-response rates among young adults, leading to under-reporting of specific mental disorders among this population (de Graaf, Have and Van Dorsselaer, 2010^[69]).

Figure 3.2. Young people, men and those with lower levels of education are less likely to respond to mental health questions

Panel A: Share of non-response rates for mental health questions (MHI-5) across different socio-demographic groups, European OECD 26, 2013



Panel B: Share of non-response rates for mental health questions (SWEWMWBS, GHQ-12 and SF-12 mental health component), United Kingdom, 2019



Note: Refer to Annex 2.B for more information about specific tools.

Source: Panel A: OECD calculations based on the 2013 *European Union Statistics on Income and Living Conditions (EU-SILC)* (n.d.^[65]), (database), <https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions>; Panel B: OECD calculations based on University of Essex, Institute for Social and Economic Research (2022^[70]), *Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009* (database). 15th Edition. UK Data Service. SN: 6614, <http://doi.org/10.5255/UKDA-SN-6614-16>, from wave 10 only (Jan 2018 – May 2020).

StatLink  <https://stat.link/1im7az>

Data on previous diagnoses for, or experienced symptoms of, specific mental disorders are likely to under-report population prevalence due to a combination of reticence to disclose personal medical history and inability to afford or access care (Hinshaw and Stier, 2008^[53]). Furthermore, prevalence of mental ill-health based on these data is heavily influenced by the characteristics of health care systems in different countries and regions, including their ability to treat and diagnose a wide range of patients. For example, data predating the pandemic show that 67% of working-age adults who wanted mental health services reported difficulty in accessing treatment (OECD, 2021^[71]). A survey in Canada compared the self-reported use of mental health services from the Canadian Community Health Survey with health service administrative data from the government of Quebec (*Régie de l'assurance maladie du Québec – RAMQ*), reporting significant discrepancies: 75% of mental health service users in the RAMQ registry did not report using services in the CCHS, with these disparities being highest for older people, those with lower levels of education and mothers of young children (Drapeau, Boyer and Diallo, 2011^[72]). Another study for Australia examined the extent of under-reporting of mental illness by matching self-reported mental health information (self-report diagnosis and self-reports of prescription drug use) to administrative records of filled prescriptions for mental disorders; the researchers found that survey respondents are significantly more likely to under-report mental illnesses compared to other health conditions because of stigma (Bharadwaj, Pai and Suziedelyte, 2017^[73]).

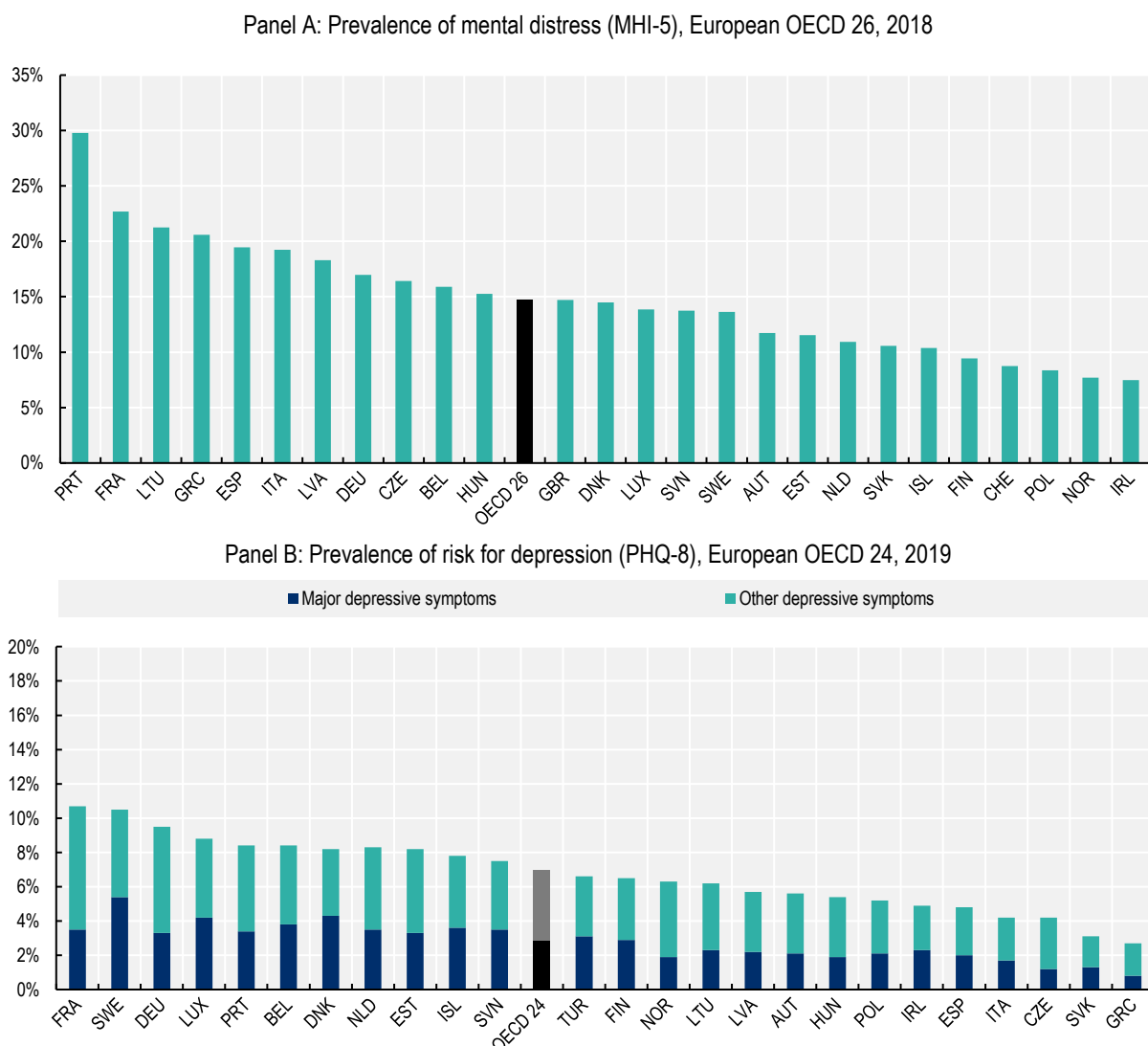
Box 3.6. Key messages: Non-response bias and missing values

- Those with worse underlying mental health may be more likely to refuse to participate in surveys, thereby understating the actual prevalence of mental ill-health; however, the evidence is not conclusive.
- There is conclusive evidence that self-reported data on previous diagnoses and experienced symptoms of specific mental ill-health conditions are significantly influenced by stigma and bias.
- Analysis from European OECD countries shows that younger people, men and those with lower levels of education are more likely to refuse to answer questions on mental health.

Are the reliability and validity of these measures consistent across cultures and socio-demographic groups?

Governments tasked with promoting population mental health need high-quality information to understand inequalities in mental health outcomes and whether national trends (either improvements or deteriorations) are masking differences within groups, so that they can target policy interventions to those who are most in need. For these reasons, a mental health indicator needs to be able to compare age cohorts, genders, race and ethnic groups, different education and income levels and other socio-economic markers. Ensuring comparability, however, is not straightforward. Cultural differences in perceptions of mental health may make some groups less likely to answer (or honestly answer) questions surrounding mental health. These challenges are true for both inter- and intra-country comparisons.⁹

Figure 3.3. Prevalence of psychological distress and depressive symptoms risk varies by as much as 100 percent across European OECD countries



Note: Panel A: risk for psychological distress is defined as having a score ≥ 52 on a scale from 0 (least distress) to 100 (most); Panel B: a respondent is deemed to be at risk for major depressive disorder if they answer “more than half the days” to either of the first two questions on the PHQ-8, and, in addition, if five or more of the eight items are reported as “more than half the days”. They are at risk for “other depressive disorders” if they answer “more than half the days” to either of the first two questions on the PHQ-8, and in addition, a total of two to four of the eight items are reported as “more than half the days” (Eurostat, n.d.^[74]). Refer to Annex 2.B for more information on individual screening tools. Source: Panel A: OECD calculations based on the 2018 *European Union Statistics on Income and Living Conditions (EU-SILC)* (n.d.^[65]), (database), <https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions>; Panel B: *European Health Interview Survey (EHIS) wave 3 data* (n.d.^[66]) (database), [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_\(EHIS\)](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_(EHIS)).

StatLink  <https://stat.link/ocxvgt>

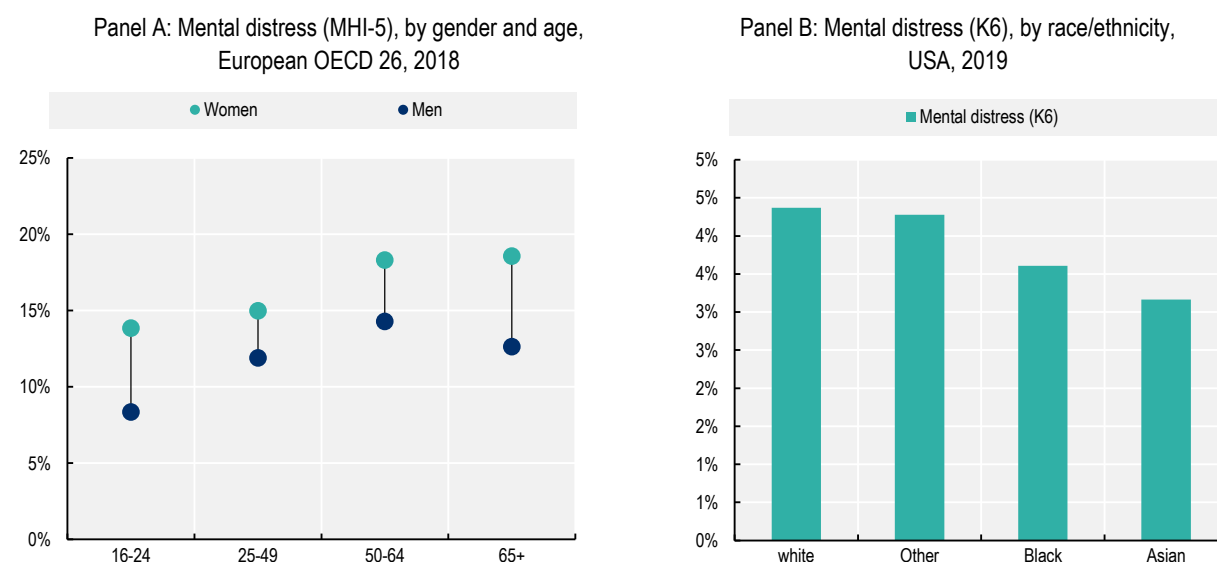
Data from European countries show large variations in the prevalence of psychological distress and depressive symptoms. The prevalence of psychological distress in Portugal, France and Lithuania is more than twice that of Ireland, Norway, Poland and Switzerland (Figure 3.3, Panel A). Similarly, the prevalence of depressive symptoms in France, Sweden and Germany is more than double that of Greece, the Slovak Republic and the Czech Republic, among others (Figure 3.3, Panel B). Yet how much of this is due to

differences in the underlying mental health of each population, and how much is due to cultural differences leading to differential response behaviours for these screening tools?

Some of these cross-country differences could stem from different levels of stigma towards mental health, with countries having lower overall prevalence levels also showing higher levels of stigma (refer to the previous section, and Table 3.2).


Comparisons of the prevalence of mental ill-health can be difficult within countries, as well. Panel A of Figure 3.4 shows that women in 26 European OECD countries are more likely to report higher levels of psychological distress than men, at all stages of their life. Panel B of Figure 3.4 also suggests that white Americans have higher levels of psychological distress than other racial/ethnic groups, and that Asian-Americans have the lowest levels. Research has shown that there are systematic gender differences in self-report bias, as men tend to minimise their symptoms more than women do (Brown et al., 2018^[63]). One study also found that men, but not women, reported fewer depressive symptoms when consent forms indicated that a more involved follow-up might occur (Sigmon et al., 2005^[75]). A survey on attitudes towards mental health and stigma in Sweden found that women were more likely to report feeling positive attitudes towards those with mental health conditions than did men (Folkhälsomyndigheten, 2022^[67]). Therefore how much of the visible difference is due to differences in reporting rather than differences in actual underlying mental health?

Figure 3.4. Are differences in reported outcomes by gender and race/ethnicity due to differences in underlying mental health or to measurement issues?



Note: Scoring information for Panel A: risk for psychological distress is defined as having a score ≥ 52 on a scale from 0 (least distress) to 100 (most); Panel B: risk for psychological distress is defined as having a score ≥ 13 on a scale from 0 (least distress) to 24 (most). Refer to Annex 2.B for more information on individual screening tools.

Source: Panel A: OECD calculations based on the 2018 *European Union Statistics on Income and Living Conditions (EU-SILC)* (n.d.^[65]), (database), <https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions>; Panel B: OECD calculations based on University of Michigan (2021^[76]), *Panel Study of Income Dynamics* (database), <https://psidonline.isr.umich.edu/default.aspx> data from 2019 only.

StatLink  <https://stat.link/3yntk5>

To answer these questions on measurement bias and cross-group comparability, researchers assess whether surveys have structural invariance by key socio-demographic characteristics in the process of clinically validating screening tools. Screening tools for symptoms of depression and anxiety, along with

the WHO-5 and SWEMWBS scales for positive mental health, are the tools that have been most frequently validated across numerous settings (e.g. gender, age cohorts and racial/ethnic groups).

In terms of screening tools for specific mental ill-health conditions, both the PHQ-8 and GAD-7 have been found to be free from basic gender and age biases. The PHQ-8 and -9 have been validated across a number of clinical settings, with different age groups, cultures and racial/ethnic groups (El-Den et al., 2018^[8]), and the PHQ-2 has been validated for use in youth and adolescents (Richardson et al., 2010^[77]). A study of the PHQ-4, limited to the United States, found no structural invariance by gender and age: its findings may extend to other countries with similar population structures, but not necessarily to others with different population structures (Sunderland et al., 2019^[33]).

Positive mental health composite scales also perform strongly on age and gender generalisability. The WHO-5 has been shown to have good construct validity for all age groups and has been deemed suitable for children aged 9 and above (Topp et al., 2015^[49]). The MHC-SF performs well across sex, age cohorts and education levels (Santini et al., 2020^[48]). WEMWBS was originally validated for an adult population but has since been validated for use in youth aged 11 and above (Warwick Medical School, 2021^[78]). In the course of validating the 14-item version of WEMWBS, researchers found evidence that two items showed bias for gender; for example, for any level of mental well-being, men were more likely than women to answer positively for the item “I’m feeling more confident” (Stewart-Brown et al., 2009^[45]). These two items were removed in the process of creating the 7-item version of the screening tool (SWEMWBS). This short form displays no response rate differences by gender, marital status or household income (Tennant et al., 2007^[14]).

Evidence for validity across racial groups is more mixed for all tools, and much of the evidence comes from either the United States or Canada. There is mixed support for cross-cultural invariance of the CES-D’s factor structure across Latino and Anglo-American populations (Crockett et al., 2005^[79]; Posner et al., 2001^[80]); one study found that item-level modifications were needed for the CES-D when administered to older Hispanic/Latino and Black respondents (El-Den et al., 2018^[8]). Other studies also indicate that Asian-American and Armenian-American populations have a different factor structure, higher depressive symptoms, and a tendency to over-endorse positive affect items, in comparison to Anglo-Americans (Iwata and Buka, 2002^[81]; Demirchyan, Petrosyan and Thompson, 2011^[82]). Research in the United Kingdom implementing the GHQ-12 across diverse racial and ethnic groups found some suggestive evidence of differences by group, requiring further study (Bowe, 2017^[83]). A study comparing Korean-American and Anglo-American older adults found that cross-cultural factors may significantly influence the diagnostic accuracy of depression scales and potentially result in the use of different cut-off scores for different populations (Lee et al., 2010^[84]). Another study revealed that Black/African-American participants with high GAD symptoms scored lower on the GAD-7 than other participants with similar GAD symptoms (Parkerson et al., 2015^[85]; Sunderland et al., 2019^[33]).

Measures of self-reported mental health may also be susceptible to bias by racial/ethnic identity. US studies have found that ethnicity appears to moderate the relationship between SRMH and a range of mental health conditions. For example, Black and Hispanic/Latino Americans are more likely to report excellent SRMH than white Americans and show a weaker association between SRMH and service use. A study in Canada found that Asian identity was associated with worse SRMH even after controlling for socio-economic status (Ahmad et al., 2014^[51]).

Many screening tools have been translated into multiple languages and used in surveys across the globe. The WHO-5, K10, MHI-5, GAD-7 and WEMWBS have all been translated into a number of languages (Sunderland et al., 2019^[33]); the WHO-5, for example, has been translated into more than 30 languages and implemented in surveys in six continents (Topp et al., 2015^[49]).¹⁰ WEMWBS has been used across 50 countries and translated into 36 languages (Stewart-Brown, 2021^[16]; Warwick Medical School, 2021^[78]). Psychometric evaluations for the MHC-SF have also been conducted in many countries (Petrillo et al., 2015^[47]; Joshanloo et al., 2013^[86]; Guo et al., 2015^[46]); however, cross-country comparisons in rates of

flourishing show a high degree of variability, some of which may be driven by measurement issues rather than differences in latent mental health (Santini et al., 2020^[48]).

Cultural differences pose significant challenges in establishing uniform definitions and descriptions of mental health and threaten cross-country comparisons (see Box 3.4). Cross-cultural validation refers to whether mental health measures that were originally generated in a given culture are applicable, meaningful and thus equivalent in another culture (Huang and Wong, 2014^[87]). Most widely used mental health scales have been developed and validated in high-income, Western and English-speaking populations (e.g. North America, Europe, Australia) and therefore assume a Western understanding of mental disorders and symptoms (Sunderland et al., 2019^[33]). This can raise questions as to their applicability to other population groups. For example, a review of 183 published studies on the mental health status of refugees reported that 78% of the findings were based on instruments that were not developed or tested specifically in refugee populations (Hollifield et al., 2002^[88]).

Evidence on cross-cultural validation for different tools is mixed. WEMWBS has been validated in 17 different languages and local populations as well as for minority populations within the United Kingdom (Warwick Medical School, 2021^[78]). Although the PHQ has been validated in many settings and is considered to be one of the more robust screening tools, one study found that, when applied in middle- or low-income countries, it performed well only in student samples and not in clinical samples, leading researchers to suggest that it should be used only in settings with relatively high rates of literacy (El-Den et al., 2018^[8]; Ali, Ryan and De Silva, 2016^[32]). Similarly, scoring schemes – i.e. the process of determining what score on a screening tool designates risk for a specific mental issue – are often calibrated based on the US general population, where the initial clinical study took place. The scoring scheme of the Kessler scales was designed to seek out maximum precision in the 90th – 99th percentile of the general population distribution, because of US epidemiological evidence that, in any given year, between 6% and 10% of the US population meet the definition of having a serious mental illness. Therefore, these scoring schemes may not be appropriate for other populations with different structures (Kessler et al., 2002^[89]). As another example, the mental health component of the SF-12 is typically scored using US-derived item weights for each response category (Ware et al., 2002^[90]). International comparisons have been done in Europe and Australia, which have found these weights to be appropriate (Vilagut et al., 2013^[91]), but this does not necessarily extend to other regions.

Research is clearly needed on culturally specific mental health scales developed using a bottom-up and open-ended approach, or with a greater degree of local adaptation, and with further testing of existing scales across different cultures and ethnicities (Sunderland et al., 2019^[33]). Furthermore, advances in psychometric models and computational statistics have led to new developments in the administration and scoring of screening tools, which can facilitate cross-cultural analyses.¹¹ Yet it is important to contextualise the magnitude of these differences. Research using data from the Gallup World Poll covering 150 countries on cross-country differences in measures of positive mental health, including life satisfaction, has found that cultural differences account for only 20% of inter-country variation in outcomes. This 20% includes both the impact of different cultures on outcomes, as well as potential measurement bias, an amount that is small in comparison to the impact of objective conditions – such as income, education and employment (Exton, Smith and Vandendriessche, 2015^[92]; OECD, 2013^[11]). The impact that these objective life conditions have on mental health is also likely to be larger than that of cultural bias. This does not negate the importance of better designing and validating mental health tools across populations, but it does provide a needed reminder that mental health indicators are informative and useful for policy.

Box 3.7. Key messages: Accuracy across groups

- Differences in attitudes toward mental health can lead to differential reporting across countries, as well as by gender, age and racial/ethnic identity.
- Surveys on stigma and discriminatory views have shown differences in attitudes toward mental health by age and gender.
- In the process of validation, screening tools are tested for biases by age, gender and racial/ethnic group. While most screening tools for specific mental ill-health conditions and most composite scales for positive mental health perform well for age and gender, evidence for race/ethnicity is mixed. More research is needed on the performance of self-reported general mental health questions.
- Survey items must be validated in local populations to ensure their suitability. Validation studies conducted in one geographic area, or in one population group, may not be applicable to other contexts.

How comparable are measures over time?

A key goal of policy makers is to understand trends over time. Is population mental health improving or deteriorating? Do policy interventions lead to visible changes in mental health outcomes? It is therefore important that the accuracy of chosen indicators hold not only cross-sectionally but also over time. There are two complications in measuring mental health over time: (1) behavioural and attitudinal changes towards mental health, leading to different response behaviour; and (2) the fact that many of the screening tools have been validated against clinical diagnoses in cross-sectional studies, which may not provide sufficient evidence that they are sensitive to changes over time.

Attitudes towards mental health have changed over the years, and while stigma and bias remain, progress in reducing them has been made. In recent years, governments across the OECD have pursued public information campaigns, especially centred in schools and educational institutions, to destigmatise mental illness. Even before the COVID-19 pandemic, 12 OECD countries waged national campaigns to improve mental health literacy, and five had regional or local campaigns (OECD, 2021^[71]). Initial evidence of the impact is mixed: while some studies show little to no decline in stigma to mental health conditions, especially in the long run (Deady et al., 2020^[93]; Walsh and Foster, 2021^[94]), others point to an increase in service use, such as visits to psychiatric emergency departments (Cheng et al., 2016^[95]). A study in the United Kingdom found that exposure to mental health campaigns may have led to an increase in these symptoms among young people; the research suggests that this was not because of a newfound awareness of pre-existing feelings, but a causal result of increased information about mental illness (Harvey, n.d.^[96]). Other early research in this vein posits that awareness campaigns may lead to individuals categorizing their feelings and emotions – which may be mild or moderate – as more concerning indications of mental distress, which may then change their own perceptions and behaviours, thus leading to actual worsening of symptoms (Foulkes and Andrews, 2023^[97]).

If anti-stigma campaigns are indeed having their intended impact, then general population attitudes toward poor mental health may be changing, and the average person may feel more comfortable speaking openly and honestly about their mental health. This could distort estimates of mental ill-health prevalence over time. If the general population ten years ago felt less comfortable honestly answering questions on how often they felt “down, depressed or hopeless” over the past two weeks, one might expect higher rates of non-response, or of respondents lying about their true feelings, than today; as a result, one would expect to see the reported prevalence of psychological distress *increase* just because of this change in attitudes.

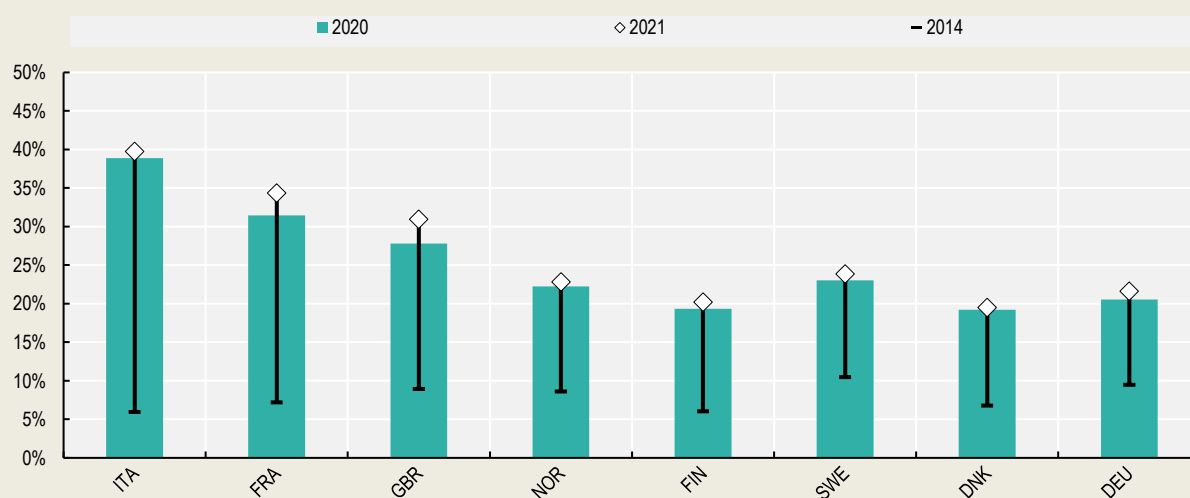
Box 3.8. Changes in mental health during the pandemic

During and in the wake of the COVID-19 pandemic, mental health deteriorated in most OECD countries, with rates of symptoms of depression and anxiety doubling in some (OECD, 2021^[98]; OECD, 2021^[57]). Indeed, for eight European OECD countries that have comparable pre-pandemic baseline data, the share of the population at risk for depression rose substantially, and by more than 20 percentage points in Italy and France (Figure 3.5; (OECD, 2021^[99])). A study looking at data from January 2020 to January 2021 estimated that the share of people experiencing symptoms of anxiety and depression were 28% and 26% higher, respectively, in 2020 than they would have been had the pandemic not occurred (OECD, 2021^[57]).¹² Both longitudinal and cross-sectional studies in European countries have found that positive mental health – measured through the WHO-5 (an affect-based measure), and SWEMWBS or the MHC-SF (combining aspects of affect, eudaimonia, social connections and life evaluation) – significantly deteriorated over the course of the pandemic (Thygesen et al., 2021^[100]; Vistisen et al., 2022^[101]; Eurofound, 2021^[102]; Vinko et al., 2022^[103]).

While the increase in prevalence of symptoms for mental ill-health is more or less agreed upon, it remains to be seen whether this increase is temporary, or whether mental health will revert to pre-pandemic levels relatively quickly. As of mid-2021, overall mental health had not recovered to pre-pandemic levels; however, there were suggestions of recovery in some OECD member states (OECD, 2021^[57]; OECD, 2021^[99]). Even still, certain population groups who were particularly negatively affected, such as young people, continue to face many challenges (OECD, 2021^[99]).

Figure 3.5. Symptoms of depression rose substantially in eight European OECD countries in the first year of the pandemic

Share of respondents at risk of depression, 2020 and 2021 vs. 2014



Note: Data from 2020 and 2021 come from a different data source than do data from 2014, meaning that caution should be taken in interpreting numerical increases in any individual country. Both data sources use the PHQ-2 as a measure for depression risk. Data for 2020 and 2021 come from the YouGov COVID-19 behaviour tracker: 2020 pooled averages run from April through December, and 2021 pooled averages run from January through June. Baseline data come from the European Health Interview Survey (EHIS) wave 2 in 2014. Refer to Annex 2.B for more information on individual screening tools.

Source: OECD (2021^[99]), *COVID-10 and Well-being: Life in the Pandemic*, OECD Publishing, Paris, <https://dx.doi.org/10.1787/1e1ecb53-en>.


StatLink  <https://stat.link/szdacx>

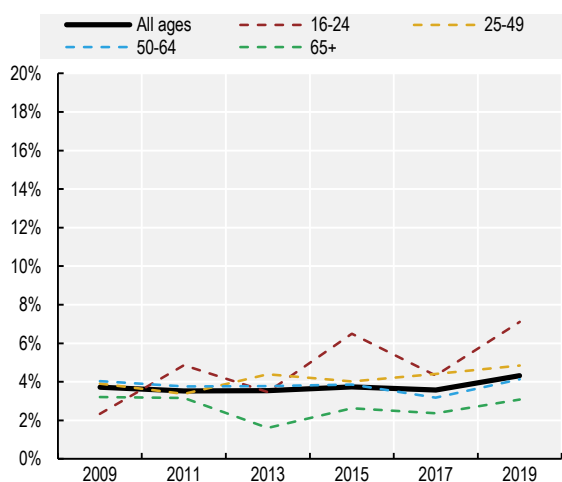
Figure 3.6 provides some evidence disproving this hypothesis. Pre-pandemic data from over 20 European OECD countries and the United States show either improving or stable mental health in the years following the financial crisis and preceding COVID-19 (prevalence of symptoms of anxiety and depression rose dramatically in 2020 at the onset of the pandemic, see Box 3.8). Prevalence of psychological distress in the United States from 2009 to 2019 (measured bi-annually using the K6 screening tool) remained broadly stable over the decade, hovering around 4% (Panel A). Although not controlling for any socio-demographic factors, this suggests that concerns surrounding changing perceptions leading to large changes in response behaviour resulting in higher prevalence rates may not hold. That said, there is some evidence of higher, and potentially rising, prevalence among young people aged 16 to 24, which may reflect a combination of changing circumstances (socio-political, economic, climate-related), changing attitudes among young people toward mental health (making them more willing to speak openly to an enumerator), and smaller sample sizes in this cohort (leading to more noise in the data). Across 26 European OECD countries, psychological distress decreased between 2013 and 2018, which would not be expected if changes in behaviours made respondents more likely to speak honestly about their poor mental health (Panel C); a similar story is shown in Panel D, which shows psychological flourishing in 24 European OECD countries rose between 2011 and 2016. Conversely, data from the United Kingdom (Panel B) show a deterioration of population mental health (as measured by the mental health component of the SF-12), while both their positive mental health (SWEMWBS) and the share at risk for a common mental disorder (GHQ-12) remained more or less stable.

One possible reason why some population surveys may show relatively stable mental health prevalence over time could be that those measures lack sensitivity to change. Many mental health screening tools were validated in cross-sectional clinical samples; researchers therefore caution that they have not been tested for sensitivity to changes over time, which can only be assessed with longitudinal data (Ahmad et al., 2014^[51]; Tennant et al., 2007^[14]; Moriarty, Zack and Kobau, 2003^[21]; Spitzer et al., 2006^[11]). However there are exceptions. Some studies have found that the PHQ-8 and PHQ-2 are sensitive to change over time (Löwe et al., 2010^[13]). In terms of positive mental health, both WEMWBS and the shorter SWEMWBS have been shown to be sensitive to change over time for both groups and individuals: researchers suggest that +/- 3 points on the WEMWBS scale, and +/- 1 to 3 points on the SWEMWBS scale, indicate a significant change (NHS Health Scotland, 2016^[15]; Shah et al., 2018^[104]).

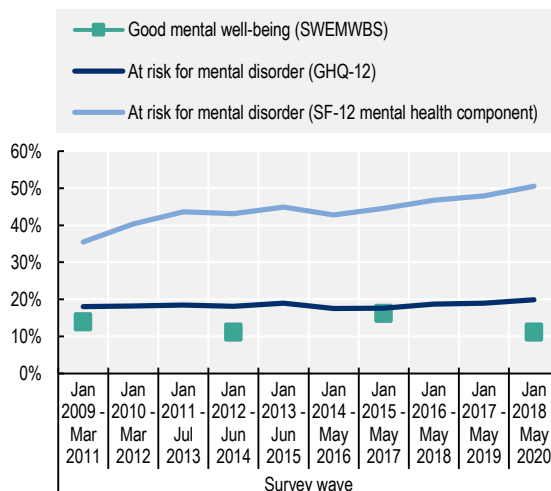
One approach to identifying measurement bias that is driven by changes in individuals' characteristics or circumstances over time is to use longitudinal data. A study by Ploubidis et al. (2019^[105]) used two nationally representative surveys in the United Kingdom to track age cohorts over two decades. Using a generalised latent variable measurement modelling framework, researchers tested whether respondents' answers to questions on the Malaise Inventory (a 9-item survey measuring psychological distress) were affected by the passage of time and found little evidence for the presence of bias in the form of age effects, survey design, period effects or cohort-specific effects.

Figure 3.6. Until 2019, mental health improved somewhat in European OECD countries, and remained roughly stable in the United States, despite greater mental health awareness

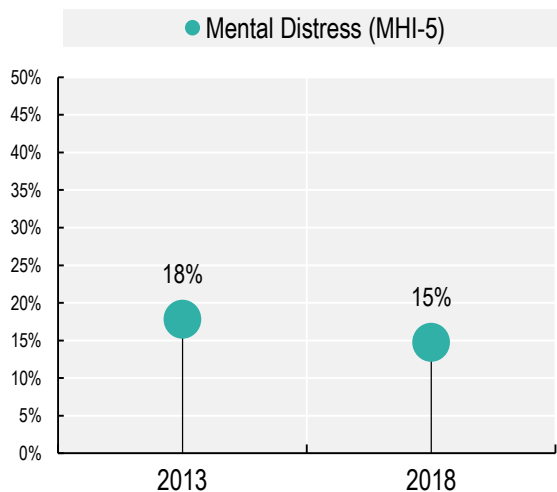
Panel A: Share of the population at risk for mental distress (K6), by age group, USA, 2009-2019



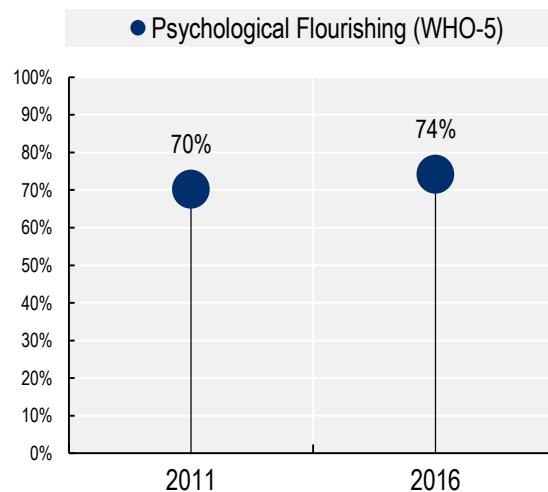
Panel B: Share of the population with good mental health (SWEMWBS) or at risk for a mental disorder (GHQ-12, and mental health component of the SF-12), GBR, 2009-2020



Panel C: Share of the population at risk for mental distress (MHI-5), European OECD 26, 2013 vs. 2018



Panel D: Share of the population who are flourishing (WHO-5), European OECD 24, 2011 vs. 2016



Note: Panel A: Scoring information for each screening tools; risk for psychological distress if score is ≥ 13 on a scale from 0 (least distress) to 24 (most). Panel B: at risk for a mental condition if score is ≤ 50 on the transformed SF-12 mental health component composite scale, 0 indicates worst mental health and 100 best possible mental health; risk for a probable common mental disorder (CMD) if score is ≥ 4 on the GHQ-12, as used in (Woodhead et al., 2012_[106]); good mental health is defined as having a SWEMWBS score more than one standard deviation above the sample average. Panel C: risk for psychological distress if score is ≥ 52 on a scale from 0 (least distress) to 100 (most); Panel D: psychological flourishing if score is ≥ 14 on a scale from 0 (worst mental health outcome) to 24 (best). Refer to Annex 2.B for more information. Source: Panel A: OECD calculations based on University of Michigan (2021_[76]), *Panel Study of Income Dynamics* (database), <https://psidonline.isr.umich.edu/default.aspx>; Panel B: OECD calculations based on University of Essex, Institute for Social and Economic Research (2022_[70]), *Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009* (database), 15th Edition. UK Data Service. SN: 6614, <http://doi.org/10.5255/UKDA-SN-6614-16>; Panel C: OECD calculations based on the 2013 and 2018 *European Union Statistics on Income and Living Conditions (EU-SILC)* (n.d._[65]), (database), <https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions>; Panel D: OECD calculations based on the 2011 and 2016 *European Quality of Life Survey* (n.d._[107]), (database), <https://www.eurofound.europa.eu/surveys/european-quality-of-life-surveys>.

Box 3.9. Key messages: Accuracy over time

- Changes in attitudes towards mental health conditions, and comfort discussing these topics openly, may lead to measurement bias when comparing prevalence rates over time.
- While evidence from impact assessments of anti-stigma campaigns is scarce, pre-pandemic evidence from cross-country trend data does not show the clear increase in reported psychological distress that lower stigma would imply.
- Some mental health screening tools – including the PHQ-8 and (S)WEMWBS – have been found to be sensitive to change over time in longitudinal studies, but other tools have not been subject to sensitivity analysis.
- More research into the presence of bias in the form of age, period or cohort-specific effects for mental health outcomes should be done.

Data collection

The practicalities of data collection can have an important impact on respondent behaviour, affecting the comfort and ease with which they interact with an enumerator and thus answer questions. Whether questions are framed as positive or negative statements, the way in which data are collected (enumerator- vs. self-administered) shapes the quality of the eventual output. Because of the sensitive nature of mental health questions, especially for screening tools that deal with suicide and suicidal ideation, additional protocols should be put in place to ensure both respondent and enumerator safety and well-being.

How does question wording affect respondents' attitudes and response behaviour?

The order in which questions are asked, and the way in which questions are framed, may prime respondents to answer in a certain way. OECD research into subjective well-being shows that the influence of question ordering on life evaluation and affect questions can be significant; because of this, subjective well-being questions should be placed early on in surveys to minimise interference from other modules (OECD, 2013^[1]).

For mental health questions, there is some evidence suggesting that questions may be upsetting to respondents, raising ethical concerns. Some studies have shown that participating in a survey with distressing questions, or answering questions focusing on distressing life events, can increase respondents' stress and worsen their mood (Labott et al., 2013^[108]), especially among populations already at risk for psychological distress. However, other research into the impacts of mental health surveys on the mood of respondents has not found evidence of significant effects (Jorm et al., 1994^[109]; Jacomb et al., 1999^[110]). The small portion of interviewees who did report feeling distress were more likely to be young women and people lacking social support (Jacomb et al., 1999^[110]).

Within a given mental health screening tool or composite scale, framing questions in a positive or negative light can impact on responses. Some tools use only negative question framing (e.g. PHQ-8, CES-D, K6), some only positive (i.e. (S)WEMWBS, WHO-5), and some employ a mix of the two (e.g. GHQ-12, MHI-5). A negatively framed question might ask, for example, how often someone felt downhearted and depressed, whereas a positively framed question might ask how often someone felt cheery and light-hearted. A respondent may feel more comfortable answering that s/he "rarely" felt cheery, rather than answering that s/he "always" felt depressed. This point is illustrated in Table 3.3, which relies on data from the UKHLS survey. The same sample of respondents were asked questions from three different mental health screening surveys. There are overlaps in the types and content of questions asked; pairs of questions are

showcased in the top portion of the table. The correlation in responses are highest for questions that appear in tools with similar tone framing, either positive or negative (e.g. feeling downhearted and depressed from the mental health component of the SF-12 vs. feeling unhappy and depressed from the GHQ-12), and lowest for items that come from tools that use different framing (e.g. been able to face up to problems from the GHQ-12 and dealing with problems well from SWEMWBS).

Users of mental health services in the United Kingdom have expressed a preference for survey tools that focus on positive, rather than negative, emotions (Stewart-Brown, 2021^[16]). A study conducted with users there found that respondents found it “upsetting” to be asked a series of negative items in mental health questionnaires, and they expressed a preference for questions – WEMWEBS, specifically – that focus on aspects of good mental health (Crawford et al., 2011^[111]).

Table 3.3. The correlation of answers to similar questions depends on whether the phrasing is positive or negative

Correlation between similarly worded questions on different mental health screening tools

Question phrasing			
	GHQ-12	SF-12	SWEMWBS
	Been able to face up to your problems		Been dealing with problems well
	Been feeling unhappy and depressed	Felt downhearted and depressed	
		Felt calm and peaceful	Been feeling relaxed
Answer correlations			
	Feeling unhappy/depressed (GHQ-12 and SF-12): 0.67	Facing up to problems (GHQ-12 and SWEMWBS): 0.36	Feeling relaxed (SF-12 and SWEMWBS): 0.56

Note: Correlations show the pairwise Pearson correlation coefficient between similarly phrased questions appearing on different mental health screening tools or scales from the same longitudinal survey. Refer to Annex 2.B for more information about specific tools.

Source: OECD calculations based on University of Essex, Institute for Social and Economic Research (2022^[70]), *Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009* (database). 15th Edition. UK Data Service. SN: 6614, <http://doi.org/10.5255/UKDA-SN-6614-16>, from wave 10 only (Jan 2018 – May 2020).

Box 3.10. Key messages: Question framing

- OECD research has shown that the ordering of subjective questions in household surveys can influence responses, therefore consistency in ordering across surveys is important, and whenever possible these questions should appear early in surveys.
- Evidence from the United Kingdom’s *Understanding Society* survey shows that whether a concept is framed in a negative or a positive light can lead to different responses.
- Some users of mental health services have expressed a preference for survey tools that focus on positive rather than negative emotions.

Does the survey mode affect respondents’ answers?

Survey modes, i.e. the way in which data are collected from respondents, can influence how respondents process and reply to questions, as well as how much information they feel comfortable revealing. One of the main drivers of differential responses based on survey mode is social desirability bias: the tendency to

present oneself in a favourable light and/or provide responses that conform to prevailing social norms. Social desirability has two components: impression management and self-deception (Paulhus, 1984^[112]). Research has shown that interview subjects under-report taboo topics and over-report socially desirable actions (Krumpal, 2013^[59]; Presser and Stinson, 1998^[113]). Social desirability bias can present itself in different ways, depending on the way in which data are collected – by an interviewer or self-administered, in person, or over the phone or Internet.

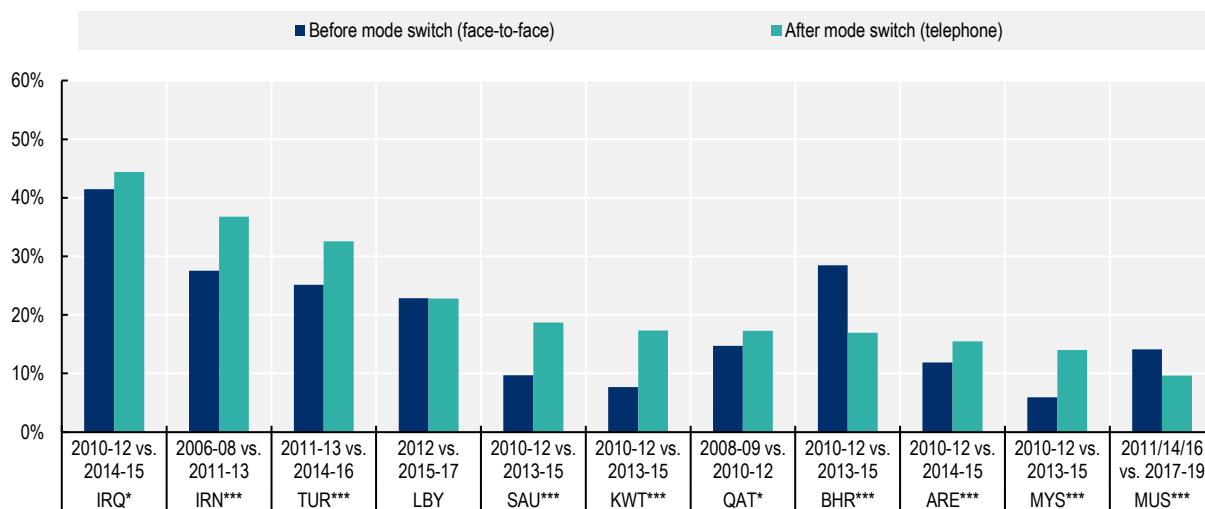
Research has found that respondents are more likely to report better physical and mental health outcomes in interviewer-administered surveys as compared to self-administered surveys. Research has also shown that self-administered survey respondents are less likely to report stigmatised medical conditions, including anxiety and mood disorders (Krumpal, 2013^[59]; Latkin et al., 2017^[114]). A study in Norway found that respondents were more likely to report symptoms of anxiety and depression in self-administered surveys as compared to interviewer-administered (either in person or over the phone) surveys (Moum, 1998^[115]); the presence of social desirability bias was particularly strong for young, well-educated respondents. Another study comparing computer-assisted self-interviewing (ACASI) with interviewer-administered paper-and-pencil (I-PAPI) surveys concluded that respondents were more likely to report mental health symptoms – as measured by the WHO-CIDI – in the self-administered survey than in the interviewer-administered one (Epstein, Barker and Kroutil, 2001^[116]), with large differences for major depressive episodes and generalised anxiety disorder.¹³

When surveys are administered by an interviewer, evidence is inconsistent as to whether respondents report better mental health outcomes face-to-face than over the phone or Internet. Evidence from the Canadian Community Health Survey (CCHS) suggests that while some physical health indicators are subject to mode effects, mental health outcomes are not¹⁴ (St-Pierre and Béland, 2004^[117]). A study in the United States found the opposite impact, with respondents exhibiting stronger social desirability behaviour in telephone interviews than in person (Holbrook, Green and Krosnick, 2003^[118]). Finally, in comparing proctored web-based assessments to paper-and-pencil administration modes, a recent meta-analytic review reported that web-based surveys do not offer an advantage regarding socially desirability in self-report questionnaires, and that the mode of administration does not affect reporting of mental health symptoms (Gnambs and Kaspar, 2017^[119]).

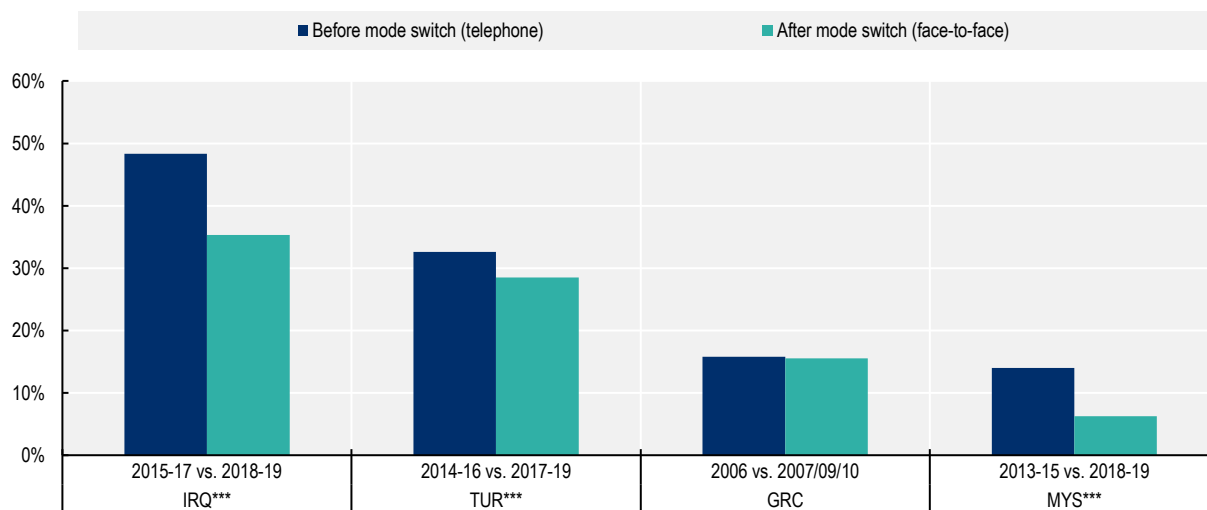
Mode effects in interviewer-administered surveys are illustrated by Figure 3.7. The figure shows the share of the population with a negative affect balance – defined as reporting to have experienced more negative, rather than positive, emotions on the previous day – from Gallup World Poll data. Gallup conducts annual surveys in over 150 countries, including all 38 OECD countries. Data are collected via telephone surveys in many OECD countries; however, face-to-face interviews are common in many places in Latin America, the Middle East, Asia, Africa and former Soviet countries.¹⁵ In a handful of countries, the survey mode changed over the past decade, switching from face-to-face to telephone survey administration and vice versa (indeed, some countries – such as Iraq, Türkiye and Malaysia – have switched multiple times). Figure 3.7 shows that negative affect balance rose (meaning worsening mental health) in eight of 11 countries after switching from in-person to telephone interviews (Panel A); similarly, negative affect balance improved in all four countries after switching from telephone to in-person (Panel B). This is in line with the findings of (Holbrook, Green and Krosnick, 2003^[118]), that respondents may be more influenced by social desirability bias when speaking to an interviewer over the phone, and thus over-report good well-being in telephone surveys.¹⁶

Figure 3.7. Shifts from in-person to telephone-administered surveys are associated with deteriorations in negative affect balance

Panel A: Share of the population with negative affect balance, three-year pooled average before and after mode switch, switch from face-to-face to telephone interviews, 11 countries, varying years




Panel B: Share of the population with negative affect balance, three-year pooled average before and after mode switch, switch from telephone to face-to-face interviews, 4 countries, varying years



Note: Countries followed by *** experienced statistically significant (at the 5% level) changes in outcomes following mode changes; * indicates the change is statistically significant at the 10% level. Three-year averages are shown (three years preceding vs. three years following a mode change); exceptions are made for countries that do not have sufficient years of data collection on either side of a mode switch. In those instances, one- or two-year averages are shown instead. IRQ and ARE did not collect negative affect data in 2013, the year of mode change.

Source: OECD calculations based on the *Gallup World Poll* (n.d._[120]) (database), <https://www.gallup.com/178667/gallup-world-poll-work.aspx>.

StatLink  <https://stat.link/ngkd9r>

Box 3.11. Key messages: Mode effects

- Respondents are more likely to report worse mental health outcomes when surveys are self-administered, as compared to interviewer-administered.
- When surveys are interviewer-administered, there is conflicting evidence as to whether mental health outcomes are subject to mode effects.
- Consistency in mode is encouraged; when the survey mode changes between data collection rounds, this should be explicitly stated.

What additional protocols or procedures should data collectors take on board for mental health modules?

Interviewer training is crucial to the quality of responses in any survey. However, the measurement of mental health raises additional issues, because of the sensitive nature of the subject matter. Although a body of trained interviewers will generally contribute to higher response rates and better responses, interviewers may struggle to garner responses to questions if they cannot explain adequately to respondents why collecting such information is important and how it will be used. In some cases, respondents may fail to understand why a public agency might want to collect this type of information. To manage risks around respondent attitudes to questions on mental health, it is imperative that interviewers are well-briefed, not just on what concepts the questions are trying to measure, but also on how the information collected will be used. This is essential in order for interviewers to build a strong rapport with respondents, which can help to improve response rates along with the quality of those responses.

A respondent's relationship with the interviewer matters. In a study conducted with a group of mental health service users in a clinical setting, respondents emphasised that a questionnaire was "only as good as the doctor who uses it" (Crawford et al., 2011^[111]). In fact, users stated that the interviewer mattered most – more than either the content or length of the survey. A study in the United States found that respondents were more likely to disclose sensitive information about illegal drug use in a face-to-face interview, as opposed to over the phone, with the difference more pronounced for Black compared to white Americans (Aquilino, 1994^[121]). A Norwegian study found minimal impact of interviewer gender and age on reported mental health symptoms, but noted that young male interviewers received fewer reports of symptoms as compared to interviewers of other ages and/or female interviewers (Moum, 1998^[115]). This can function in the opposite direction as well, with a strong interviewer-respondent bond leading to more information being disclosed.¹⁷

Recent research has shed more light on the need to involve those with lived experience in the survey design and data collection process, by building a pipeline of researchers with psychiatric disabilities and/or lived experience of mental health conditions (Jones et al., 2021^[122]; Banfield et al., 2018^[123]; Hancock et al., 2012^[124]). There is a strong basis of evidence showing that peer-interviewing techniques – drawing enumerators from the same community as interviewees – can be an effective way of improving trust between interviewer and interviewee, helping to collect high-quality data for hard-to-reach population groups (Dewa et al., 2021^[125]; Warr, Mann and Tacticos, 2011^[126]; Hancock et al., 2012^[124]). Furthermore, in the mental health context, the involvement in research of those with lived mental health experience can improve both the credibility of findings and the likelihood of their adoption into policy (Scholz et al., 2021^[127]).

Questions on suicide or suicidal ideation require careful consideration and well-designed procedures to provide needed support to respondents who are at risk (Lakeman and FitzGerald, 2009^[128]). The final item of the PHQ-9 asks about suicidal thoughts and ideation, and for this reason it is often excluded from

population surveys. In both the European Health Interview Survey (EHIS) and the United States' National Health Interview Survey (NHIS), the PHQ-8 is used instead, for precisely this reason.

For countries that do include questions on the topic of suicide, additional protocols are often employed. For example, the Australian non-suicidal self-injury (NSSI) prevalence study dealt with the sensitive nature of the survey questions by sharing in advance a large amount of information with the households to be interviewed; this helped to alleviate ethical considerations, and did lead to lower non-response rates (Taylor et al., 2011^[129]). In implementing the Mental Health and Access to Care Survey (MHACS), the Canadian government provides mental health resources to both respondents *and* interviewers; enumerators are also provided with employee support services to help them navigate stress or ill effects to their own mental health that may be induced by administering the questionnaire (response to an OECD questionnaire, 2022).

Box 3.12. Key messages: Interviewer training

- Respondents are more likely to participate in an interview, and answer truthfully, if they feel comfortable with the interviewer. Enumerator training should focus on building rapport and trust with respondents.
- Careful procedures and support practices must be in place if surveys are to include questions surrounding suicide and suicidal ideation; in the case of household surveys, it may be best practice to avoid these types of questions on ethical grounds.

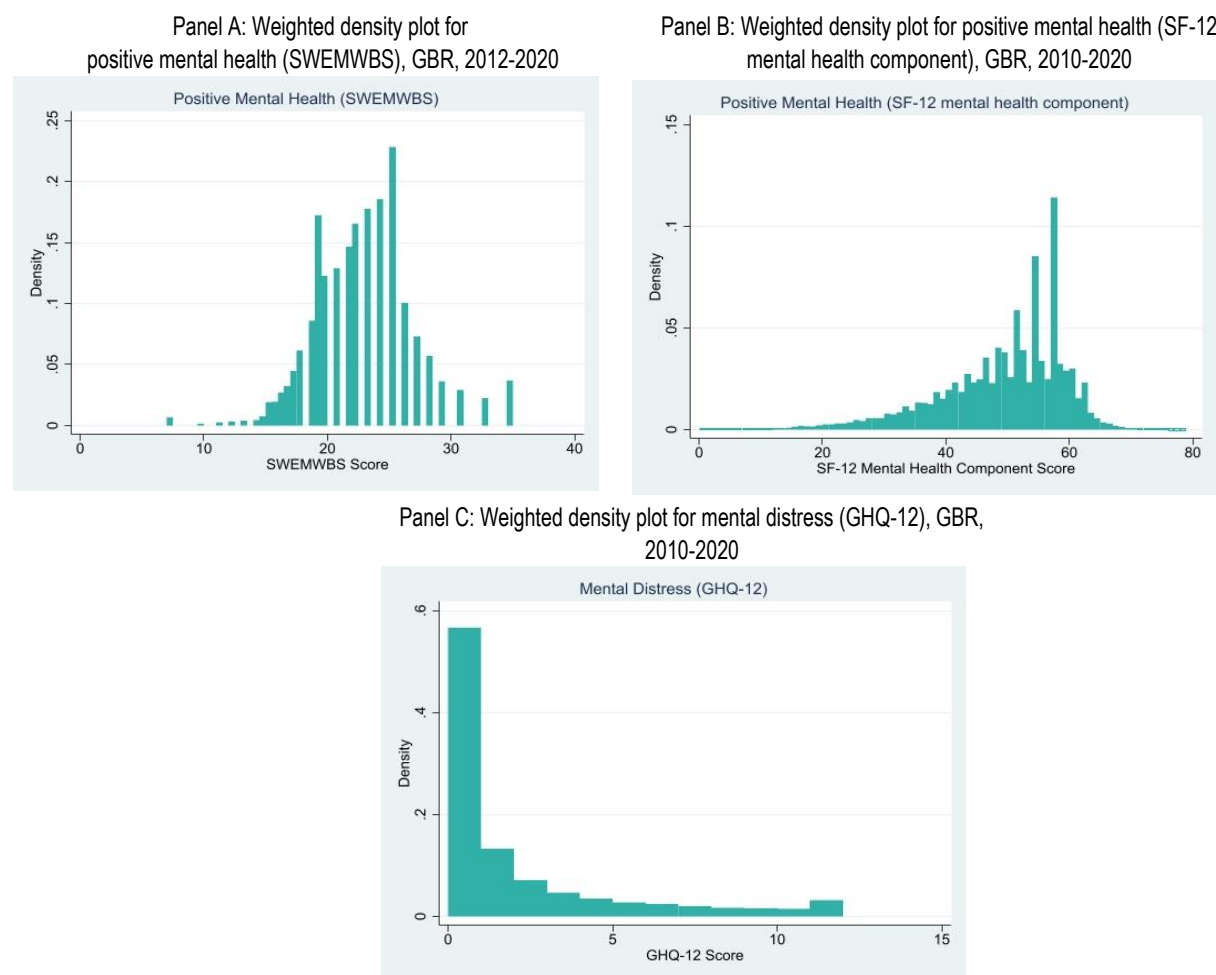
Analysis

Once mental health data are collected, as accurately and consistently as possible, they are only useful for policy makers when used in analysis. Many of the data described in this report are not binary, meaning the outcome variable is measured on a scale. This is true for the screening tools and composite scales, which contain a number of items, coded and scored accordingly, ranging from worst to best possible mental health. General mental health status tools are typically single questions; however, answer options are not binary and typically use a Likert scale (the exact number and phrasing of answer options varies across scales, see Table 2.11).

What are the trade-offs between using cut-off scores vs. continuous measures?

Having a continuous outcome measure for mental health has many benefits, including that it provides more detailed and nuanced information about population mental health. Further, when the distribution of responses is normal, without floor or ceiling effects,¹⁸ this allows for parametric analysis of outcomes, which enables researchers to better analyse the impacts of given policies or interventions. For example, research into screening tools for positive mental health has shown that data sourced from (S)WEMWBS have a distribution that more closely approximates a normal distribution than do screening tools that focus on specific mental illnesses (Shah et al., 2021^[17]).¹⁹ Indeed, this can be seen visually in Figure 3.8, which shows density plots for the three mental health tools included in the tenth wave of the UKHLS survey: SWEMWBS; the mental health component of the SF-12 (MHC-12); and the GHQ-12. Of the three, SWEMWBS most closely approximates a normal distribution, followed by the SF-12, while the GHQ-12 shows significant floor effects. Separate research into the MHI-5 has shown that it is positively skewed; it is better able to distinguish between those with worse mental health than between those with higher levels of positive mental health (Elovanio et al., 2020^[6]; Thorsen et al., 2013^[130]).

Figure 3.8. Positive mental health scales may better approximate a normal distribution



Note: Density plots showing the weighted scores for: Panel A: SWEMWBS (ranging from 9.5 as worse mental well-being, and 35 as better mental well-being); Panel B: the mental health components of the SF-12 (ranging from 0, low functioning, to 100, high functioning); and Panel C: the GHQ-12 (ranging from 0, better mental health, to 12, worse mental health) is lowest. Data for SF-12 and the GHQ-12 are from waves 2 to 10 of *Understanding Society*; data for SWEMWBS come from waves 4, 7 and 10. Refer to Annex 2.B for more information on individual screening tools.

Source: OECD calculations based on University of Essex, Institute for Social and Economic Research (2022^[70]), *Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009* (database). 15th Edition. UK Data Service. SN: 6614, <http://doi.org/10.5255/UKDA-SN-6614-16>, from wave 10 only (Jan 2018 – May 2020).

While normal distributions are useful for regression analysis, in order for mental health information to be useful at either a micro-level (i.e. primary care physician, conducting a screening interview to see if a patient is at risk and requires more support) or macro-level (i.e. a government office tasked with tracking changes in risk over time), it is often useful to use cut-off scores to group outcomes into categories. These categories vary depending on the screening tool and scoring convention used, but typically encompass things such as “at risk for depression”, “at risk for anxiety”, “major depressive disorder”, “severe psychological distress”, “psychological flourishing”, etc. These categories can also be useful in analysis to understand how mental health interacts with other aspects of well-being: for example, the share of the employed or unemployed who are experiencing anxiety, or the quality of social connections for those at risk for depression compared to those who are not (OECD, 2021^[57]).

One general criticism of the use of thresholds is that they can be arbitrary. However, in the case of mental health screeners, thresholds are established through a rigorous validation process; researchers use receiver operating characteristic (ROC) analysis to determine which cut-off score maximises both the sensitivity and the specificity of the measure (see Box 3.3).²⁰ Cut-off scores are also useful in that they convert responses to a series of screening tool questions into something comparable to the results of an in-depth diagnostic interview: risk for a certain mental health condition. The PHQ, GAD, Kessler and CES-D surveys all have standard, validated cut-off scores (Kessler et al., 2002^[89]; Kroenke et al., 2007^[44]; Moriarty, Zack and Kobau, 2003^[21]; Manea, Gilbody and McMillan, 2015^[131]; Kroenke et al., 2009^[12]; Spitzer et al., 2006^[111]). The traditional CES-D cut-off score indicative of “depressive case” in clinical samples is 16, but this threshold has been known to produce a high rate of false positives in non-clinical samples (Eaton, 2004^[132]; Santor and Coyne, 1997^[133]). The GAD-7 also has established cut-off scores, but studies have found that it performs better at identifying the share of the population at risk for generalised anxiety disorder and less well at picking up on other types of anxiety disorders, such as social anxiety disorder (Beard and Björgvinsson, 2014^[134]; Sunderland et al., 2019^[33]).

Though the GHQ-12 is commonly used to screen general mental health conditions, it has been found to generate a high level of false positives; one study found that as many as half of those identified as having a mental disorder were false positives (positive predictive value of 0.53) (Schmitz, Kruse and Tress, 2001^[4]). Other mental ill-health screening tools like the MHI-5 or GHQ-12 were not developed with a standard validated cut-off to define a case of common mental disorder. Although these scales may not have an internationally comparable cut-off score, they have been validated in several studies. For instance, (Berwick et al., 1991^[38]) validated the MHI-5 as a measure for depression using clinical interviews as the gold standard and reported an optimal cut-off score of 52.²¹ Subsequent research has corroborated the finding that the MHI-5 performs well as a screener for depression and general mood disorders but much less well as a measure for anxiety, somatoform disorders and substance use disorders (Rumpf et al., 2001^[135]; Strand et al., 2003^[7]; Thorsen et al., 2013^[130]).

Some tools have multiple accepted cut-off scores, depending on the intended diagnosis, meaning varying scoring conventions can lead to different prevalence estimates. Figure 3.9 shows the density plot for PHQ-8 scores, ranging from 0 (least at risk for depression) to 25 (most at risk) for 22 European OECD countries. The vertical lines show different validated thresholds. A score of 10 or above indicates risk for major depressive disorder (shown in black) (Kroenke et al., 2008^[136]). Other threshold categorisations deem a score of 5-9 as risk for mild depression, 10-14 as moderate, 15-19 as risk for depression, and 20+ as risk for severe depression (Kroenke, Spitzer and Williams, 2001^[137]). Another scoring convention (not shown in Figure 3.9), used by Eurostat, is not based purely on the raw score but rather defines major depressive symptoms by respondent answers to individual questions.²² All three measures lead to different prevalence estimates from the same underlying dataset: (1) 6.9% at risk for major depressive disorder; (2) 15.2% at risk for mild depression, 2.9% at risk for moderate, 1.7% at risk for moderately severe and 0.8% at risk for severe; and (3) 3.1% with major depressive symptoms.²³

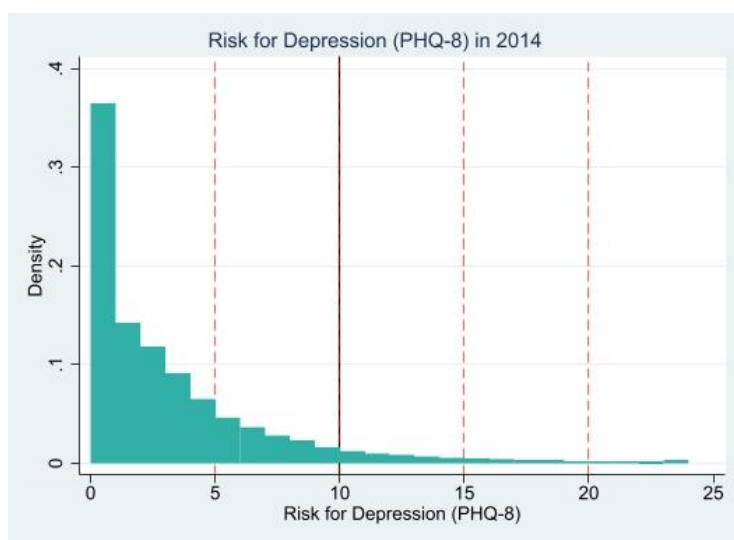
Although there is no clinical gold standard for psychological well-being, positive mental health composite scales have also developed cut-off scores at the request of users. Two main approaches have been put forward for (S)WEMWBS: one statistical and the other benchmarking.²⁴ In the first, researchers recommend cut-off points at +/- one standard deviation, which result in approximately 15% of the population having high well-being and 15% having low well-being. In the second approach, cut-off scores for (S)WEMWBS are benchmarked against measures capturing symptoms of depression and anxiety (see below for a more detailed discussion of positive mental health tools being used as screeners for mood disorders). Studies have benchmarked WEMWBS against the CES-D, PHQ-9 and GAD-7 to suggest cut-off points on the WEMWBS scale that indicate risk for probable clinical depression, possible depression or mild depression (and/or anxiety). Taking all of this together, researchers suggested that a cut-off point of 60 on the WEMWBS scale, and of 28 on the SWEMWBS scale, can be used to identify the top 15% of

those with high mental well-being, but caution that because there is no clinical measure of high mental well-being these thresholds are by definition arbitrary (Warwick Medical School, 2021^[78]).

The MHC-SF comprises three subscales (emotional, social and psychological well-being), which can be scored to group individuals into three categories: flourishing, languishing, and for those in neither of the previous two categories, moderately mentally healthy (Lamers et al., 2011^[20]). However surveys in Canada and Denmark have found that the majority of the population scores highly enough to be categorized as flourishing, which runs counter to the theory that flourishing and languishing represent a minority of the population and are deviations from the average. This suggests that more conservative scoring criteria could be warranted to improve the sensitivity of the measure (Santini et al., 2020^[48]).

Figure 3.9. Different scoring conventions can lead to different estimates of prevalence

Density plots showing distribution of risk for depression (PHQ-8), European OECD 22, 2014



Note: Weighted density plot for PHQ-8 scores in 22 European OECD countries; scores range from 0 (lowest risk) to 25 (highest risk for depression). Vertical lines indicate validated cut-off scores as established in the literature: As shown by the bold black vertical line, a score ≥ 10 or above indicates risk for major depressive disorder (Kroenke et al., 2008^[136]); as shown by the dotted red line vertical lines, a score of 5-9 as risk for mild depression, 10-14 as moderate, 15-19 as risk for depression, and 20+ as risk for severe depression (Kroenke, Spitzer and Williams, 2001^[137]).

Source: OECD calculations based on *European Health Interview Survey* (EHIS) wave 2 data (n.d.^[66]) (database), [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_\(EHIS\)](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_(EHIS)).

Box 3.13. Key messages: Continuous measures vs. cut-off scores

- Mental health tools that provide a continuous, as opposed to binary, outcome variable provide a more nuanced view of population mental health.
- Evidence suggests that positive mental health composite scales better approximate a normal distribution than do measures of psychological distress.
- Cut-off scores provide researchers and policy makers with clear categories of who is at risk and who is not. While cut-off scores for mental ill-health screening tools have been validated against clinical diagnoses to maximise their sensitivity and specificity, no such gold standard exists for psychological flourishing.

- Despite best efforts to ensure sensitivity and specificity, cut-off scores may provide false positives or be ill-suited for some population groups.
- Different scoring conventions for the same screening tool can lead to different prevalence estimates, therefore care should be taken to ensure consistency.

Conclusion

The questions addressed in this chapter are important to consider when thinking about which tools are best for measuring population mental health. As with any survey data, a number of challenges exist, and care is needed when interpreting changes over time and across groups. Perhaps more unique to mental health, stigma and discriminatory views can contribute to bias in reported data. Furthermore, it is important to integrate the perspective of those with lived experience in survey design to ensure the quality and policy relevance of data. However, the evidence reviewed in this chapter shows that existing mental health tools provide useful and policy-relevant outcomes. Given the increasing urgency of the mental health crisis and the prioritisation of action on this front by governments, collecting high-quality mental health data following existing good practice is all the more important. On-going research into open questions of measurement can then progress in tandem with the continual monitoring of population mental health.

All OECD countries are currently measuring population mental health in one form or another and are already making cross-group and longitudinal comparisons. While additional research is needed to test the sensitivity of some tools to change over time, the high-frequency data collected during the COVID-19 pandemic showed that many mental health measures are indeed sensitive to change. Whereas policy discussions prior to COVID-19 sometimes emphasised that rates of common mental health conditions like generalised anxiety disorder and depressive disorders had remained stable in recent years, there is now broad consensus that the pandemic caused a dramatic increase in rates of psychological distress over the first two years – and these spikes have been captured in the data collected in OECD countries.

The task ahead is to better harmonise data collection and provide recommendations for quality improvement for initiatives already underway. The results of a 2022 OECD questionnaire, showcased in Chapter 2, illustrate remaining gaps in the type of mental health outcomes collected by countries: an absence of a harmonised approach to measure symptoms of anxiety, a lack of standardisation in affective and eudaimonic tools, and very uneven use of tools that measure non-depression, non-anxiety types of specific mental health conditions. The recommendations for these areas made below take into account the practical considerations of data collectors, noting the need to keep any new survey items short.

Based on a comparative assessment of the statistical quality of different tools, their response burden and cost (proxied by item length) as well as information on existing data collection practices (Table 3.1), **the report recommends the inclusion of specific mental health outcome measurement tools for national statistical offices to adopt in household, social and health surveys.** These recommendations do not imply the phasing out of other tools that OECD countries are already using to capture population mental health outcomes, particularly with regard to previous diagnoses and experienced symptoms, or measures from the 2013 *OECD Guidelines on Measuring Subjective Well-being* (such as life satisfaction). Rather, they offer a small set of instruments on which a more internationally harmonised set of population mental health outcome indicators could be built:

- *Mental ill-health –priority recommendation:* The Patient Health Questionnaire-4 (**PHQ-4**) could be included in more frequent surveys, alongside the regular collection of the PHQ-8 or PHQ-9 in health surveys. The PHQ-4 measure combines two depression questions from the longer PHQ-9 scale and two anxiety questions from the GAD-7 screening tool. It covers both depression and anxiety, rather than focusing on only one of these two most common mental health conditions. Furthermore, it does so with only four questions, keeping the module relatively short and with a low response

burden. 81% of OECD countries are already implementing the PHQ-8 or PHQ-9, meaning there is trend data to which the depression questions in the PHQ-4 could be linked.²⁵ The PHQ-8/9 and the GAD-7 could be retained in specific health surveys, while the PHQ-4 could be introduced in general, more frequent surveys, given its shorter length.

- *Positive mental health –recommendation:* Either the **WHO-5** or **SWEMWBS** could be used to measure affective and eudaimonic aspects of positive mental health in a standardised way across countries. These suggestions are mainly based on trends in country measurement practice. The **WHO-5** is a tool for measuring positive affect in that it is relatively short and easy to implement, is included in the OECD’s Subjective Well-being Guidelines as an experimental affect module (OECD, 2013_[1]), has been translated into many languages, and has been found to be reliable and valid. Although currently used by only 16% of OECD countries, it has been recommended for use by other OECD projects, including as a part of a conceptual framework for measuring the non-financial performance of firms (Siegerink, Shinwell and Žarnic, 2022_[138]) and in an effort to use patient-reported indicator surveys (PaRIS) to centre health care delivery on the outcomes that matter to patients (de Bienassis et al., 2021_[139]). **SWEMWBS** is a more comprehensive tool, in that it covers affective, eudaimonic, and social connections aspects of positive mental health. This makes it slightly longer than the WHO-5, though only by two questions. SWEMWBS – or the longer 14-question WEMWBS – has been adopted by 19% of OECD countries. For countries already active in subjective well-being or positive mental health measurement, some of the indicator items within SWEMWBS may overlap with existing data collection efforts to measure concepts such as life evaluation and the quantity and quality of social connections (see (OECD, 2020_[140]) and (OECD, 2013_[1]) for existing OECD recommendations and examples). In these instances, the WHO-5 may be more suitable in that it covers only affect. The topic of measuring affect and eudaimonia specifically will continue to be explored in future OECD workstreams on subjective well-being.
- *General mental health status – recommendation:* A single question about a respondent’s **general mental health status** could be included in a range of different surveys across a country’s entire data infrastructure system. Single general mental health questions have less of an evidence base compared to established screening tools, but the findings that do exist suggest it is a useful and meaningful measure. Many OECD countries already collect data on self-reported physical health, thus in question framing it will be important to distinguish between self-reported *physical* vs. self-reported *mental* health. Some OECD countries currently collect a general self-reported health measure that captures both physical and mental health; we recommend separating these measures out. In order for this to happen in an internationally comparable way, more research and coordination must happen to align existing country efforts. Canada has been an early adopter of this approach, and its framing as a self-reported mental health (SRMH) question-and-answer option has already been adopted by Chile and Germany; furthermore, much of the existing evidence-base on these types of question has been produced in Canada. Other countries interested in adding such an item to surveys may be interested in using this framing as well: “In general, how is your mental health? Excellent / Very good / Good / Fair / Poor.”

Currently, very few countries are using tools to collect information on mental health conditions beyond depression and anxiety, such as substance use disorders, PTSD, obsessive compulsive disorder, eating disorders, bipolar disorder, etc. There are exceptions – these outcomes are covered by all countries that use structured interviews (see Table 2.4), and France and Slovenia, among a few others, have implemented detailed survey modules with tools that capture diagnoses and symptoms of these conditions. There is value in measuring these concepts as distinct conditions, rather than as a part of general mental ill-health, therefore the statistical agenda moving forward could focus on developing recommendations in this space.

As a general point to note, it is more informative, and thus a better use of limited resources, to diversify *across* tool types and mental health outcome measures, rather than to implement a variety of iterations of the same type of tool. For example, rather than implementing a range of different screening tools to capture depression/anxiety across a country's survey infrastructure, it would be of greater use to harmonise *within* tool areas. This might mean choosing a single depression/anxiety screening tool, then supplement it with single-item question tools to capture received diagnoses, experience of symptoms and so on.

Above all, this report has highlighted the importance of precision when communicating outcome measures. Each tool measures a specific, slightly different facet of population mental health. Furthermore, individual tools can be scored in a variety of ways, each leading to different estimates for mental health outcomes. This speaks to the need for greater harmonisation, but also of clearer communication in terms of stating what is meant by mental health and how it is measured. This is all the more important given the rise of mental health to the top of national agendas in the years following the pandemic.

References

- Ahmad, F. et al. (2014), "Single item measures of self-rated mental health: A scoping review", *BMC Health Services Research*, Vol. 14/398, pp. 1-11, <https://doi.org/10.1186/1472-6963-14-398>. [51]
- Ahn, J., Y. Kim and K. Choi (2019), "The psychometric properties and clinical utility of the Korean version of GAD-7 and GAD-2", *Frontiers in Psychiatry*, Vol. 12/10, <https://doi.org/10.3389/fpsy.2019.00127>. [10]
- Ali, G., G. Ryan and M. De Silva (2016), "Validated screening tools for common mental disorders in low and middle income countries: A systematic review", *PLoS ONE*, Vol. 11/6, <https://doi.org/10.1371/journal.pone.0156939>. [32]
- Aquilino, W. (1994), "Interview mode effects in surveys of drug and alcohol use: A field experiment", *Public Opinion Quarterly*, Vol. 58/2, pp. 210-240, <https://doi.org/10.1086/269419>. [121]
- Banfield, M. et al. (2018), "Lived experience researchers partnering with consumers and carers to improve mental health research: Reflections from an Australian initiative", *International Journal of Mental Health Nursing*, Vol. 27/4, pp. 1219-1229, <https://doi.org/10.1111/INM.12482>. [123]
- Beard, C. and T. Björgvinsson (2014), "Beyond generalized anxiety disorder: Psychometric properties of the GAD-7 in a heterogeneous psychiatric sample", *Journal of Anxiety Disorders*, Vol. 28/6, pp. 547-552, <https://doi.org/10.1016/j.janxdis.2014.06.002>. [134]
- Berwick, D. et al. (1991), "Performance of a five-item mental health screening test", *Medical Care*, Vol. 29/2, pp. 169-176, <https://doi.org/10.1097/00005650-199102000-00008>. [38]
- Bharadwaj, P., M. Pai and A. Suziedelyte (2017), "Mental health stigma", *Economics Letters*, Vol. 159, pp. 57-60, <https://doi.org/10.1016/j.econlet.2017.06.028>. [73]
- Böhnke, J. and T. Croudace (2016), "Calibrating well-being, quality of life and common mental disorder items: Psychometric epidemiology in public mental health research", *The British Journal of Psychiatry*, Vol. 209/2, pp. 162-168, <https://doi.org/10.1192/BJP.BP.115.165530>. [50]
- Botha, F., P. Butterworth and R. Wilkins (2021), "Taking the pulse of the nation: Validating a single-item measure of mental distress", *Melbourne Institute Working Paper Series*, No. 6/21, Melbourne Institute of Applied Economic and Social Research, The University of Melbourne, <https://melbourneinstitute.unimelb.edu.au/publications/working-papers/search/result?paper=3821299>. [52]
- Bowe, A. (2017), "The cultural fairness of the 12-item general health questionnaire among diverse adolescents", *Psychological Assessment*, Vol. 29/1, pp. 87-97, <https://doi.org/10.1037/PAS0000323>. [83]
- Brown, S. et al. (2018), "Mental health and reporting bias: Analysis of the GHQ-12", *IZA Discussion Paper* no. 11771, <https://www.iza.org/publications/dp/11771/mental-health-and-reporting-bias-analysis-of-the-ghq-12>. [63]

- Cheng, J. et al. (2016), "Impact of a mass media mental health campaign on psychiatric emergency department visits", *Canadian Journal of Public Health*, Vol. 107/3, [95]
<https://www.jstor.org/stable/90006480>.
- Coles, M. and S. Coleman (2010), "Barriers to treatment seeking for anxiety disorders: Initial data on the role of mental health literacy", *Depression and Anxiety*, Vol. 27/1, pp. 63-71, [56]
<https://doi.org/10.1002/DA.20620>.
- Cornelius, B. et al. (2013), "The performance of the K10, K6 and GHQ-12 to screen for present state DSM-IV disorders among disability claimants", *BMC Public Health*, Vol. 13/128, [37]
<http://www.biomedcentral.com/1471-2458/13/128>.
- Crawford, M. et al. (2011), "Selecting outcome measures in mental health: The views of service users", *Journal of Mental Health*, Vol. 20/4, pp. 336-346, [111]
<https://doi.org/10.3109/09638237.2011.577114>.
- Crockett, L. et al. (2005), "Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: A national study", *Journal of Consulting and Clinical Psychology*, Vol. 73/1, pp. 47-58, [79]
<https://doi.org/10.1037/0022-006X.73.1.47>.
- Dadfar, M. et al. (2018), "Reliability, validity, and factorial structure of the World Health Organization-5 Well-Being Index (WHO-5) in Iranian psychiatric outpatients", *Trends in Psychiatry and Psychotherapy*, Vol. 40/2, pp. 79-84, [18]
<https://doi.org/10.1590/2237-6089-2017-0044>.
- de Bienassis, K. et al. (2021), "Patient-reported indicators in mental health care: Towards international standards among members of the OECD", *International Journal for Quality in Health Care*, Vol. 6/34(Suppl 1), [139]
<https://doi.org/10.1093/INTQHC/MZAB020>.
- de Graaf, R. et al. (2000), "Psychiatric and sociodemographic predictors of attrition in a longitudinal study the Netherlands Mental Health Survey and Incidence Study (NEMESIS)", *American Journal of Epidemiology*, Vol. 152/11, pp. 1039-1047, [60]
<https://doi.org/10.1093/AJE/152.11.1039>.
- de Graaf, R., M. Have and S. Van Dorsselaer (2010), "The Netherlands Mental Health Survey and Incidence Study-2 (NEMESIS-2): Design and methods", *International Journal of Methods in Psychiatric Research*, Vol. 19/3, pp. 125-141, [69]
<https://doi.org/10.1002/MPR.317>.
- Deady, M. et al. (2020), "Suicide awareness campaigns: Are they a valid prevention strategy?", in *What Can Be Done to Decrease suicidal Behaviour in Australia? A Call to Action*, [93]
https://www.blackdoginstitute.org.au/wp-content/uploads/2020/09/What-Can-Be-Done-To-Decrease-Suicide_Chapter-3-Awareness-Campaigns.pdf.
- Demirchyan, A., V. Petrosyan and M. Thompson (2011), "Psychometric value of the Center for Epidemiologic Studies Depression (CES-D) scale for screening of depressive symptoms in Armenian population", *Journal of Affective Disorders*, Vol. 133/3, pp. 489-498, [82]
<https://doi.org/10.1016/j.jad.2011.04.042>.
- Dere, J. et al. (2015), "Cross-cultural examination of measurement invariance of the Beck Depression Inventory-II", *Psychological Assessment*, Vol. 27/1, pp. 68-81, [149]
<https://doi.org/10.1037/pas0000026>.

- Dewa, L. et al. (2021), "Reflections, impact and recommendations of a co-produced qualitative study with young people who have experience of mental health difficulties", *Health Expectations*, Vol. 24/S1, pp. 134-146, <https://doi.org/10.1111/HEX.13088>. [125]
- Dhingra, S. et al. (2011), "PHQ-8 days: A measurement option for DSM-5 Major Depressive Disorder (MDD) severity", *Population Health Metrics*, Vol. 9/11, <https://doi.org/10.1186/1478-7954-9-11>. [42]
- Dowling, N. et al. (2016), "Measurement and control of bias in patient reported outcomes using multidimensional item response theory", *BMC Medical Research Methodology*, Vol. 16/63, <https://doi.org/10.1186/s12874-016-0161-z>. [150]
- Drapeau, A., R. Boyer and F. Diallo (2011), "Discrepancies between survey and administrative data on the use of mental health services in the general population: Findings from a study conducted in Québec", *BMC Public Health*, Vol. 11/837, pp. 1-10, <https://doi.org/10.1186/1471-2458-11-837>. [72]
- Dunn, K. et al. (2009), "Quantification and examination of depression-related mental health literacy", *Journal of Evaluation in Clinical Practice*, Vol. 15/4, pp. 650-653, <https://doi.org/10.1111/J.1365-2753.2008.01067.X>. [55]
- Easton, S. et al. (2017), "The Kessler psychological distress scale: Translation and validation of an Arabic version", *Health and Quality of Life Outcomes*, Vol. 15/1, <https://doi.org/10.1186/S12955-017-0783-9>. [9]
- Eaton, W. (2004), "Center for Epidemiologic Studies Depression scale: Review and revision (CESD and CESD-R)", in M. E. Maruish (ed.), *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment: Instruments for Adults*, Lawrence Erlbaum Associates Publishers, <https://psycnet.apa.org/record/2004-14941-011>. [132]
- Eaton, W. et al. (1992), "Psychopathology and attrition in the epidemiologic catchment area surveys", *American Journal of Epidemiology*, Vol. 135/9, pp. 1051-1059, <https://doi.org/10.1093/OXFORDJOURNALS.AJE.A116399>. [61]
- El-Den, S. et al. (2018), "The psychometric properties of depression screening tools in primary healthcare settings: A systematic review", *Journal of Affective Disorders*, Vol. 225/1, pp. 503-522, <https://doi.org/10.1016/j.jad.2017.08.060>. [8]
- Elovainio, M. et al. (2020), "General Health Questionnaire (GHQ-12), Beck Depression Inventory (BDI-6), and Mental Health Index (MHI-5): Psychometric and predictive properties in a Finnish population-based sample", *Psychiatry Research*, Vol. 289, p. 112973, <https://doi.org/10.1016/j.psychres.2020.112973>. [6]
- Epstein, J., P. Barker and L. Kroutil (2001), "Mode effects in self-reported mental health data", *Public Opinion Quarterly*, Vol. 65/4, pp. 529-549, <https://doi.org/10.1086/323577>. [116]
- Eurofound (2021), *Living, Working and COVID-19: Mental health and trust decline across EU as pandemic enters another year*, Publications Office of the European Union, <https://www.eurofound.europa.eu/topic/covid-19>. [102]
- Eurofound (n.d.), *European Quality of Life Surveys (EQLS) (database)*, <https://www.eurofound.europa.eu/surveys/european-quality-of-life-surveys> (accessed on 10 June 2022). [107]

- Eurostat (n.d.), *European health interview survey (EHIS)*, [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_\(EHIS\)](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_health_interview_survey_(EHIS)) (accessed on 16 June 2021). [66]
- Eurostat (n.d.), *European Health Interview Survey (EHIS): Reference Metadata in Euro SDMX Metadata Structure (ESMS)*, https://ec.europa.eu/eurostat/cache/metadata/en/hlth_det_esms.htm (accessed on 25 March 2022). [74]
- Eurostat (n.d.), *European Union Statistics on Income and Living Conditions (EU-SILC) (database)*, <https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions> (accessed on 10 June 2022). [65]
- Exton, C., C. Smith and D. Vandendriessche (2015), “Comparing happiness across the world: Does culture matter?”, *OECD Statistics Working Papers*, No. 2015/4, OECD Publishing, Paris, <https://doi.org/10.1787/5jrqqpzd9bs2-en>. [92]
- Folkhälsomyndigheten (2022), *Syner på psykisk ohälsa och suicid: En befolkningsundersökning om kunskaper och attityder [The view of mental illness and suicide: A population survey on knowledge and attitudes]*, <https://www.folkhalsomyndigheten.se/publikationer-och-material/publikationsarkiv/s/syner-pa-psykisk-ohalsa-och-suicid-/?pub=105538#105580>. [67]
- Foulkes, L. and J. Andrews (2023), “Are mental health awareness efforts contributing to the rise in reported mental health problems? A call to test the prevalence inflation hypothesis”, *New Ideas in Psychology*, Vol. 69, p. 101010, <https://doi.org/10.1016/J.NEWIDEAPSYCH.2023.101010>. [97]
- Fried, E. (2017), “The 52 symptoms of major depression: Lack of content overlap among seven common depression scales”, *Journal of Affective Disorders*, Vol. 208, pp. 191-197, <https://doi.org/10.1016/j.jad.2016.10.019>. [146]
- Furukawa, T. and D. Goldberg (1999), “Cultural invariance of likelihood ratios for the General Health Questionnaire”, *Lancet*, Vol. 353/9152, pp. 561-562, [https://doi.org/10.1016/S0140-6736\(98\)05470-1](https://doi.org/10.1016/S0140-6736(98)05470-1). [145]
- Furukawa, T. et al. (2001), “Stratum-specific likelihood ratios of two versions of the General Health Questionnaire”, *Psychological Medicine*, Vol. 31/3, pp. 519-529, <https://doi.org/10.1017/s0033291701003713>. [144]
- Furukawa, T. et al. (2003), “The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being”, *Psychological Medicine*, Vol. 33/2, pp. 357-362, <https://doi.org/10.1017/S0033291702006700>. [36]
- Gallup (n.d.), *Gallup World Poll (database)*, <https://www.gallup.com/analytics/318875/global-research.aspx> (accessed on 18 June 2021). [120]
- García, M. et al. (2005), “Comparison between telephone and self-administration of Short Form Health Survey Questionnaire (SF-36).”, *Gaceta sanitaria / S.E.S.P.A.S.*, Vol. 19/6, pp. 433-439, [https://doi.org/10.1016/S0213-9111\(05\)71393-5](https://doi.org/10.1016/S0213-9111(05)71393-5). [143]
- Garland, A. et al. (2018), “Use of the WHO’s perceived Well-Being Index (WHO-5) as an efficient and potentially valid screen for depression in a low income country”, *Families, Systems and Health*, Vol. 36/2, pp. 148-158, <https://doi.org/10.1037/FSH0000344>. [19]

- Gibbons, C. and S. Skevington (2018), “Adjusting for cross-cultural differences in computer-adaptive tests of quality of life”, *Quality of Life Research*, Vol. 27/4, pp. 1027-1039, <https://doi.org/10.1007/s11136-017-1738-7>. [141]
- Gill, S. et al. (2007), “Validity of the mental health component scale of the 12-item Short-Form Health Survey (MCS-12) as measure of common mental disorders in the general population”, *Psychiatry Research*, Vol. 152/1, pp. 63-71, <https://doi.org/10.1016/j.psychres.2006.11.005>. [26]
- Gnambs, T. and K. Kaspar (2017), “Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis”, *Assessment*, Vol. 24/6, pp. 746-762, <https://doi.org/10.1177/1073191115624547>. [119]
- Goldberg, D., T. Oldehinkel and J. Ormel (1998), “Why GHQ threshold varies from one place to another”, *Psychol Med*, Vol. 28/4, pp. 915-921, <https://doi.org/10.1017/s0033291798006874>. [151]
- Guo, C. et al. (2015), “Psychometric evaluation of the Mental Health Continuum-Short Form (MHC-SF) in Chinese adolescents: A methodological study”, *Health and Quality of Life Outcomes*, Vol. 13/198, <https://doi.org/10.1186/S12955-015-0394-2>. [46]
- Hancock, N. et al. (2012), “Participation of mental health consumers in research: Training addressed and reliability assessed”, *Australian Occupational Therapy Journal*, Vol. 59/3, pp. 218-224, <https://doi.org/10.1111/J.1440-1630.2012.01011.X>. [124]
- Hanmer, J., R. Hays and D. Fryback (2007), “Mode of administration is important in US national estimates of health-related quality of life”, *Medical Care*, Vol. 45/12, pp. 1171-1179, <https://doi.org/10.1097/MLR.0b013e3181354828>. [142]
- Hapke, U. et al. (2022), “Depressive symptoms in the general population before and in the first year of the COVID-19 pandemic: Results of the GEDA 2019/2020 study”, *Journal of Health Monitoring*, Vol. 7/4, <https://doi.org/10.25646/10664>. [158]
- Haroz, E. et al. (2017), “How is depression experienced around the world? A systematic review of qualitative literature”, *Social Science & Medicine*, Vol. 183, pp. 151-162, <https://doi.org/10.1016/j.socscimed.2016.12.030>. [31]
- Harvey, S. (n.d.), *Do mental health awareness campaigns work? Let's look at the evidence*, Black Dog Institute, <https://www.blackdoginstitute.org.au/news/do-mental-health-awareness-campaigns-work-lets-look-at-the-evidence/> (accessed on 25 March 2022). [96]
- Hinshaw, S. and A. Stier (2008), “Stigma as related to mental disorders”, *Annual Review of Clinical Psychology*, Vol. 4, pp. 367-393, <https://doi.org/10.1146/annurev.clinpsy.4.022007.141245>. [53]
- Hoeymans, N. et al. (2004), “Measuring mental health of the Dutch population: A comparison of the GHQ-12 and the MHI-5”, *Health and Quality of Life Outcomes*, Vol. 2/23, <https://doi.org/10.1186/1477-7525-2-2>. [154]
- Holbrook, A., M. Green and J. Krosnick (2003), “Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias”, *Public Opinion Quarterly*, Vol. 67/1, pp. 79-125, <https://doi.org/10.1086/346010>. [118]
- Hollifield, M. et al. (2002), “Measuring trauma and health status in refugees: A critical review”, *JAMA*, Vol. 288/5, pp. 611-621, <https://doi.org/10.1001/jama.288.5.611>. [88]

- Huang, F. et al. (2006), "Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients", *Journal of General Internal Medicine*, Vol. 21/6, pp. 547-552, <https://doi.org/10.1111/j.1525-1497.2006.00409.x>. [28]
- Huang, W. and S. Wong (2014), "Cross-Cultural Validation", in *Encyclopedia of Quality of Life and Well-Being Research*, Springer Netherlands, https://doi.org/10.1007/978-94-007-0753-5_630. [87]
- Iwata, N. and S. Buka (2002), "Race/ethnicity and depressive symptoms: A cross-cultural/ethnic comparison among university students in East Asia, North and South America", *Social Science and Medicine*, Vol. 55/12, pp. 2243-2252, [https://doi.org/10.1016/S0277-9536\(02\)00003-5](https://doi.org/10.1016/S0277-9536(02)00003-5). [81]
- Jacomb, P. et al. (2002), "Predictors of refusal to participate: A longitudinal health survey of the elderly in Australia", *BMC Public Health*, Vol. 2/4, <https://doi.org/10.1186/1471-2458-2-4>. [68]
- Jacomb, P. et al. (1999), "Emotional response of participants to a mental health survey", *Social Psychiatry and Psychiatric Epidemiology*, Vol. 34/2, pp. 80-84, <https://doi.org/10.1007/S001270050115>. [110]
- Jin, K. and W. Wang (2014), "Generalized IRT models for extreme response style", *Educational and Psychological Measurement*, Vol. 74/1, pp. 116-138, <https://doi.org/10.1177/0013164413498876>. [152]
- Jones, N. et al. (2021), "Lived experience, research leadership, and the transformation of mental health services: Building a researcher pipeline", *Psychiatric Services*, Vol. 72/5, pp. 591-593, <https://doi.org/10.1176/appi.ps.202000468>. [122]
- Jorm, A. et al. (1994), "Do mental health surveys disturb? Further evidence", *Psychological Medicine*, Vol. 24/1, pp. 233-237, <https://doi.org/10.1017/S0033291700026994>. [109]
- Joshanloo, M. et al. (2013), "Measurement invariance of the Mental Health Continuum-Short Form (MHC-SF) across three cultural groups", *Personality and Individual Differences*, Vol. 55/7, pp. 755-759, <https://doi.org/10.1016/J.PAID.2013.06.002>. [86]
- Kessler, R. et al. (2002), "Short screening scales to monitor population prevalences and trends in non-specific psychological distress", *Psychological Medicine*, Vol. 32/6, pp. 959-976, <https://doi.org/10.1017/S0033291702006074>. [89]
- Kessler, R. et al. (2004), "The US National Comorbidity Survey Replication (NCS-R): Design and field procedures", *International Journal of methods in Psychiatric Research*, Vol. 13/2, pp. 69-92, <https://doi.org/10.1002/MPR.167>. [64]
- Knudsen, A. et al. (2021), "Prevalence of mental disorders, suicidal ideation and suicides in the general population before and during the COVID-19 pandemic in Norway: A population-based repeated cross-sectional analysis", *The Lancet Regional Health Europe*, Vol. 4, <https://doi.org/10.1016/J.LANEPE.2021.100071>. [160]
- Kroenke, K., R. Spitzer and J. Williams (2001), "The PHQ-9: Validity of a brief depression severity measure", *Journal of General Internal Medicine*, Vol. 16/9, <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>. [137]
- Kroenke, K. et al. (2009), "An ultra-brief screening scale for anxiety and depression: The PHQ-4", *Psychosomatics*, Vol. 50/6, pp. 613-621, <https://doi.org/10.1176/appi.psy.50.6.613>. [12]

- Kroenke, K. et al. (2007), "Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection", *Annals of Internal Medicine*, Vol. 146/5, pp. 317-325, <https://doi.org/10.7326/0003-4819-146-5-200703060-00004>. [44]
- Kroenke, K. et al. (2008), "The PHQ-8 as a measure of current depression in the general population", *Journal of Affective Disorders*, Vol. 114/1-3, pp. 163-173, <https://doi.org/10.1016/j.jad.2008.06.026>. [136]
- Krumpal, I. (2013), "Determinants of social desirability bias in sensitive surveys: A literature review", *Quality and Quantity*, Vol. 47/4, pp. 2025-2047, <https://doi.org/10.1007/S11135-011-9640-9>. [59]
- Labott, S. et al. (2013), "Emotional risks to respondents in survey research: Some empirical evidence", *Journal of Empirical Research on Human Research Ethics*, Vol. 8/4, p. 53, <https://doi.org/10.1525/JER.2013.8.4.53>. [108]
- Lakeman, R. and M. FitzGerald (2009), "The ethics of suicide research", *Crisis*, Vol. 30/1, pp. 13-19, <https://doi.org/10.1027/0227-5910.30.1.13>. [128]
- Lamers, S. et al. (2011), "Evaluating the psychometric properties of the Mental Health Continuum-Short Form (MHC-SF)", *Journal of Clinical Psychology*, Vol. 67/1, pp. 99-110, <https://doi.org/10.1002/jclp.20741>. [20]
- Latkin, C. et al. (2017), "The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland", *Addictive Behaviors*, Vol. 73, pp. 133-136, <https://doi.org/10.1016/J.ADDBEH.2017.05.005>. [114]
- Lee, J. et al. (2010), "Cross-cultural considerations in administering the Center for Epidemiologic Studies Depression Scale", *Gerontology*, Vol. 57/5, pp. 455-461, <https://doi.org/10.1159/000318030>. [84]
- Leong, F., P. Priscilla Lui and Z. Kalibatseva (2019), "Multicultural Issues in Clinical Psychological Assessment", in Selbom, M. and J. Suhr (eds.), *The Cambridge Handbook of Clinical Assessment and Diagnosis*, Cambridge University Press, <https://doi.org/10.1017/9781108235433>. [23]
- Löwe, B. et al. (2010), "A 4-item measure of depression and anxiety: Validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population", *Journal of Affective Disorders*, Vol. 122/1-2, pp. 86-95, <https://doi.org/10.1016/j.jad.2009.06.019>. [13]
- Lowthian, P. and L. Lloyd (2020), *How face-to-face interviewer attitudes and beliefs moderate the effect of monetary incentive on UK Labour Force Survey response rates*, Office for National Statistics (ONS), <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/methodologies/howfacetofaceinterviewerattitudesandbeliefsmoderatetheeffectofmonetaryincentiveonuklabourforcesurveyresponserates>. [162]
- Manea, L., S. Gilbody and D. McMillan (2015), "A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression", *General Hospital Psychiatry*, Vol. 37/1, pp. 67-75, <https://doi.org/10.1016/j.genhosppsych.2014.09.009>. [131]

- Mauz, E. et al. (2022), "Time trends of mental health indicators in Germany's adult population before and during the COVID-19 pandemic", *medRxiv*, [159]
<https://doi.org/10.1101/2022.10.09.22280826>.
- Medina-Mora, M. et al. (2008), "The Mexican National Comorbidity Survey (M-NCS): Overview and results", in Kessler, R. and T. Üstün (eds.), *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*, Cambridge University Press, [153]
<https://digitallibrary.un.org/record/700329?ln=en>.
- Moriarty, D., M. Zack and R. Kobau (2003), "The Centers for Disease Control and Prevention's healthy days measures: Population tracking of perceived physical and mental health over time", *Health and Quality of Life Outcomes*, Vol. 1/37, [21]
<https://doi.org/10.1186/1477-7525-1-37>.
- Mostafa, T. et al. (2021), "Missing at random assumption made more plausible: Evidence from the 1958 British birth cohort", *Journal of Clinical Epidemiology*, Vol. 136, pp. 44-54, [62]
<https://doi.org/10.1016/J.JCLINEPI.2021.02.019>.
- Moum, T. (1998), "Mode of administration and interviewer effects in self-reported symptoms of anxiety and depression", *Social Indicators Research*, Vol. 45/1-3, pp. 279-318, [115]
<https://doi.org/10.1023/a:1006958100504>.
- Murphy, R. and B. Hallahan (2016), "Differences between DSM-IV and DSM-5 as applied to general adult psychiatry", *Irish Journal of Psychological Medicine*, Vol. 33/3, pp. 135-141, [30]
<https://doi.org/10.1017/IPM.2015.54>.
- New Zealand Government (2018), *He Ara Oranga : Report of the Government Inquiry into Mental Health and Addiction*, Mental Health and Addiction Inquiry, [34]
<https://mentalhealth.inquiry.govt.nz/inquiry-report/he-ara-oranga/>.
- NHS Health Scotland (2016), *Warwick-Edinburgh Mental Well-being Scale (WEMWBS): User guide-Version 2*, NHS Health Scotland, Edinburgh, [15]
<https://s3.amazonaws.com/helpscout.net/docs/assets/5f97128852faff0016af3a34/attachment/s/5fe10a9eb624c71b7985b8f3/WEMWBS-Scale.pdf>.
- NHS Health Scotland (2008), *Review of scales of positive mental health validated for use with adults in the UK: Technical report*, NHS Health Scotland, Edinburgh, [3]
<http://www.healthscotland.scot/media/2244/review-of-scales-of-positive-mental-health-validated-for-use-with-adults-in-the-uk.pdf>.
- O'Connor, D. and R. Parslow (2010), "Mental health scales and psychiatric diagnoses: Responses to GHQ-12, K-10 and CIDI across the lifespan", *Journal of Affective Disorders*, Vol. 121/3, pp. 263-267, [27]
<https://doi.org/10.1016/j.jad.2009.06.038>.
- OECD (2021), *A New Benchmark for Mental Health Systems: Tackling the Social and Economic Costs of Mental Ill-Health*, OECD Health Policy Studies, OECD Publishing, Paris, [71]
<https://doi.org/10.1787/4ed890f6-en>.
- OECD (2021), *COVID-19 and Well-being: Life in the Pandemic*, OECD Publishing, Paris, [99]
<https://doi.org/10.1787/1e1ecb53-en>.
- OECD (2021), *Fitter Minds, Fitter Jobs: From Awareness to Change in Integrated Mental Health, Skills and Work Policies*, Mental Health and Work, OECD Publishing, [57]
<https://doi.org/10.1787/a0815d0f-en>.

- OECD (2021), "Tackling the mental health impact of the COVID-19 crisis: An integrated, whole-of-society response", *OECD Policy Responses to Coronavirus (COVID-19)*, <https://www.oecd.org/coronavirus/policy-responses/tackling-the-mental-health-impact-of-the-covid-19-crisis-an-integrated-whole-of-society-response-0cca0b/>. [98]
- OECD (2020), *How's Life? 2020: Measuring Well-being*, OECD Publishing, Paris, <https://doi.org/10.1787/9870c393-en>. [140]
- OECD (2017), *OECD Guidelines on Measuring Trust*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264278219-en>. [2]
- OECD (2013), *OECD Guidelines on Measuring Subjective Well-being*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264191655-en>. [1]
- Ohno, S. et al. (2017), "Smallest detectable change and test-retest reliability of a self-reported outcome measure: Results of the Center for Epidemiologic Studies Depression Scale, General Self-Efficacy Scale, and 12-item General Health Questionnaire", *Journal of Evaluation in Clinical Practice*, Vol. 23/6, pp. 1348-1354, <https://doi.org/10.1111/JEP.12795>. [5]
- Parkerson, H. et al. (2015), "Cultural-based biases of the GAD-7", *Journal of Anxiety Disorders*, Vol. 31, pp. 38-42, <https://doi.org/10.1016/j.janxdis.2015.01.005>. [85]
- Paulhus, D. (1984), "Two-component models of socially desirable responding", *Journal of Personality and Social Psychology*, Vol. 46/3, pp. 598-609, <https://psycnet.apa.org/doi/10.1037/0022-3514.46.3.598>. [112]
- Pedrelli, P. et al. (2013), "Reliability and validity of the Symptoms of Depression Questionnaire (SDQ)", *CNS Spectrums*, Vol. 19/6, pp. 535-546, <https://doi.org/10.1017/S1092852914000406>. [147]
- Petrillo, G. et al. (2015), "The Mental Health Continuum–Short Form (MHC–SF) as a measure of well-being in the Italian context", *Social Indicators Research*, Vol. 121, pp. 291-312, <https://doi.org/10.1007/s11205-014-0629-3>. [47]
- Pettersson, A. et al. (2015), "Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review", *Nordic Journal of Psychiatry*, Vol. 69/7, pp. 497-508, <https://doi.org/10.3109/08039488.2015.1008568>. [41]
- Pintea, S. and R. Moldovan (2009), "The Receiver-Operating Characteristic (ROC) analysis: Fundamentals and applications in clinical psychology", *Journal of Cognitive and Behavioral Psychotherapies*, Vol. 9/1, pp. 49-66, <https://psycnet.apa.org/record/2009-04815-004> (accessed on 0). [24]
- Ploubidis, G., E. McElroy and H. Moreira (2019), "A longitudinal examination of the measurement equivalence of mental health assessments in two British birth cohorts", *Longitudinal and Life Course Studies*, Vol. 10/4, pp. 471-489, <https://doi.org/10.1332/175795919X15683588979486>. [105]
- Posner, S. et al. (2001), "Factor variability of the Center for Epidemiological Studies Depression Scale (CES-D) among urban Latinos", *Ethnicity and Health*, Vol. 6/2, pp. 137-144, <https://doi.org/10.1080/13557850120068469>. [80]

- Presser, S. and L. Stinson (1998), "Data collection mode and social desirability bias in self-reported religious attendance", *American Sociological Review*, Vol. 63/1, pp. 137-145, <https://doi.org/10.2307/2657486>. [113]
- Public Health Agency of Sweden (2022), *Ny nationell strategi för psykisk hälsa och suicidprevention*, <https://www.folkhalsomyndigheten.se/livsvillkor-levnadsvanor/psykisk-halsa-och-suicidprevention/psykisk-halsa/nationell-strategi/>. [35]
- Richardson, L. et al. (2010), "Evaluation of the Patient Health Questionnaire-9 item for detecting major depression among adolescents", *Pediatrics*, Vol. 126/6, pp. 1117-1123, <https://doi.org/10.1542/peds.2010-0852>. [77]
- Rivera-Riquelme, M., J. Piqueras and P. Cuijpers (2019), "The revised Mental Health Inventory-5 (MHI-5) as an ultra-brief screening measure of bidimensional mental health in children and adolescents", *Psychiatry Research*, Vol. 274, pp. 247-253, <https://doi.org/10.1016/J.PSYCHRES.2019.02.045>. [40]
- Rose, M. and J. Devine (2014), "Assessment of patient-reported symptoms of anxiety", *Dialogues in Clinical Neuroscience*, Vol. 16/2, pp. 197-211, <https://doi.org/10.31887/DCNS.2014.16.2/mrose>. [43]
- Rumpf, H. et al. (2001), "Screening for mental health: Validity of the MHI-5 using DSM-IV Axis I psychiatric disorders as gold standard", *Psychiatry Research*, Vol. 105/3, pp. 243-253, [https://doi.org/10.1016/S0165-1781\(01\)00329-8](https://doi.org/10.1016/S0165-1781(01)00329-8). [135]
- Santini, Z. et al. (2022), "Higher levels of mental wellbeing predict lower risk of common mental disorders in the Danish general population", *Mental Health & Prevention*, Vol. 26, p. 200233, <https://doi.org/10.1016/j.mhp.2022.200233>. [148]
- Santini, Z. et al. (2020), "Measuring positive mental health and flourishing in Denmark: Validation of the mental health continuum-short form (MHC-SF) and cross-cultural comparison across three countries", *Health and Quality of Life Outcomes*, Vol. 18/297, <https://doi.org/10.1186/s12955-020-01546-2>. [48]
- Santomauro, D. et al. (2021), "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic", *The Lancet*, Vol. 398/10312, pp. 1700-1712, [https://doi.org/10.1016/S0140-6736\(21\)02143-7](https://doi.org/10.1016/S0140-6736(21)02143-7). [161]
- Santor, D. and J. Coyne (1997), "Shortening the CES-D to improve its ability to detect cases of depression", *Psychological Assessment*, Vol. 9/3, pp. 233-243, <https://doi.org/10.1037/1040-3590.9.3.233>. [133]
- Schmitz, N., J. Kruse and W. Tress (2001), "Improving screening for mental disorders in the primary care setting by combining the GHQ-12 and SCL-90-R subscales", *Comprehensive Psychiatry*, Vol. 42/2, pp. 166-173, <https://doi.org/10.1053/COMP.2001.19751>. [4]
- Scholz, B. et al. (2021), "'People just need to try it to be converted!': A picture of consumer mental health research in Australia and New Zealand", *Issues in Mental Health Nursing*, Vol. 42/3, pp. 249-255, <https://doi.org/10.1080/01612840.2020.1795763>. [127]
- Shah, N. et al. (2021), "Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS): Performance in a clinical sample in relation to PHQ-9 and GAD-7", *Health and Quality of Life Outcomes*, Vol. 19/260, <https://doi.org/10.1186/s12955-021-01882-x>. [17]

- Shah, N. et al. (2018), "Responsiveness of the Short Warwick Edinburgh Mental Well-Being Scale (SWEMWBS): Evaluation a clinical sample", *Health and Quality of Life Outcomes*, Vol. 16/1, <https://doi.org/10.1186/s12955-018-1060-2>. [104]
- Siegerink, V., M. Shinwell and Ž. Žarnic (2022), "Measuring the non-financial performance of firms through the lens of the OECD Well-being Framework: A common measurement framework for "Scope 1" Social performance", *OECD Papers on Well-being and Inequalities*, No. 3, OECD Publishing, Paris, https://www.oecd-ilibrary.org/social-issues-migration-health/measuring-the-non-financial-performance-of-firms-through-the-lens-of-the-oecd-well-being-framework_28850c7f-en (accessed on 18 May 2022). [138]
- Sigmon, S. et al. (2005), "Gender differences in self-reports of depression: The response bias hypothesis revisited", *Sex Roles*, Vol. 53/5-6, pp. 401-411, <https://doi.org/10.1007/s11199-005-6762-3>. [75]
- Singer, E., H. Hippler and N. Schwarz (1992), "Confidentiality assurances in surveys: Reassurance or threat?", *International Journal of Public Opinion Research*, Vol. 4/3, pp. 256-268, <https://doi.org/10.1093/IJPOR/4.3.256>. [156]
- Singer, E., D. Von Thurn and E. Miller (1995), "Confidentiality assurances and response: A quantitative review of the experimental literature", *Public Opinion Quarterly*, Vol. 59/1, pp. 66-77, <https://doi.org/10.1086/269458>. [58]
- Smith, D. and B. Oreskes (2019), "Homeless population's mental illness, substance abuse under-reported", *Los Angeles Times*, <https://www.latimes.com/california/story/2019-10-07/homeless-population-mental-illness-disability>. [157]
- Spitzer, R. et al. (2006), "A brief measure for assessing generalized anxiety disorder: The GAD-7", *Archives of Internal Medicine*, Vol. 166/10, pp. 1092-1097, <https://doi.org/10.1001/ARCHINTE.166.10.1092>. [11]
- Statistics Canada (2021), *Measuring Population Mental Health in Canada*, <https://www.slideshare.net/StatsCommunications/oecd-wellbeing-and-mental-health-conference-jennifer-ali-statcan>. [29]
- Stewart-Brown, S. (2021), *15 years on: Insights and reflections on the Warwick-Edinburgh Mental Wellbeing Scales (WEMWBS)*, What Works Wellbeing, <https://whatworkswellbeing.org/resources/insights-and-reflections-on-the-warwick-edinburgh-mental-wellbeing-scales-wemwbs/>. [16]
- Stewart-Brown, S. et al. (2009), "Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey", *Health and Quality of Life Outcomes*, Vol. 7/1, p. 15, <https://doi.org/10.1186/1477-7525-7-15>. [45]
- St-Pierre, M. and Y. Béland (2004), "Mode effects in the Canadian Community Health Survey: A comparison of CAPI and CATI", in *2004 Proceedings of the American Statistical Association Meeting, Survey Research Methods*, American Statistical Association, Toronto, Canada, https://www.statcan.gc.ca/en/statistical-programs/document/3226_D16_T9_V5. [117]
- Strand, B. et al. (2003), "Measuring the mental health status of the Norwegian population: A comparison of the instruments SCL-25, SCL-10, SCL-5 and MHI-5 (SF-36)", *Nordic Journal of Psychiatry*, Vol. 57/2, pp. 113-118, <https://doi.org/10.1080/08039480310000932>. [7]

- Streiner, D. and J. Cairney (2007), "What's under the ROC? An introduction to Receiver Operating Characteristics curves", *The Canadian Journal of Psychiatry*, Vol. 52/2, pp. 121-128, <https://doi.org/10.1177/070674370705200210> (accessed on 18 May 2022). [25]
- Suhr, D. (2006), *Exploratory or Confirmatory Factor Analysis?*, SAS Users Group International (SUGI), San Francisco, CA, <https://support.sas.com/resources/papers/proceedings/proceedings/sugi31/200-31.pdf>. [22]
- Sunderland, M. et al. (2019), "Self-Report Scales for Common Mental Disorders", in *The Cambridge Handbook of Clinical Assessment and Diagnosis*, Cambridge University Press, <https://doi.org/10.1017/9781108235433.019>. [33]
- Tambling, R., C. D'Aniello and B. Russell (2021), "Mental health literacy: A critical target for narrowing racial disparities in behavioral health", *International Journal of Mental Health and Addiction*, <https://doi.org/10.1007/s11469-021-00694-w>. [54]
- Taylor, A. et al. (2011), "Methodological issues associated with collecting sensitive information over the telephone: Experience from an Australian non-suicidal self-injury (NSSI) prevalence study", *BMC Medical Research Methodology*, Vol. 11, <https://doi.org/10.1186/1471-2288-11-20>. [129]
- Tennant, R. et al. (2007), "The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Development and UK validation", *Health and Quality of Life Outcomes*, Vol. 5/63, <https://doi.org/10.1186/1477-7525-5-63>. [14]
- Thorsen, S. et al. (2013), "The predictive value of mental health for long-term sickness absence: the Major Depression Inventory (MDI) and the Mental Health Inventory (MHI-5) compared", *BMC Medical Research Methodology*, Vol. 13/115, <https://doi.org/10.1186/1471-2288-13-115>. [130]
- Thygesen, L. et al. (2021), "Decreasing mental well-being during the COVID-19 pandemic: A longitudinal study among Danes before and during the pandemic", *Journal of Psychiatric Research*, Vol. 144, pp. 151-157, <https://doi.org/10.1016/J.JPSYCHIRES.2021.09.035>. [100]
- Topp, C. et al. (2015), "The WHO-5 well-being index: A systematic review of the literature", *Psychotherapy and Psychosomatics*, Vol. 84/3, pp. 167-176, <https://doi.org/10.1159/000376585>. [49]
- University of Essex, I. (2022), *Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009 (database)*, 5th Edition. UK Data Service., <https://doi.org/10.5255/UKDA-SN-6614-16> (accessed on 10 June 2022). [70]
- University of Michigan (2021), *Panel Study of Income Dynamics (database)*, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, <https://psidonline.isr.umich.edu/default.aspx> (accessed on 10 June 2022). [76]
- Vaughn-Coaxum, R., P. Mair and J. Weisz (2016), "Racial/ethnic differences in youth depression indicators", *Clinical Psychological Science*, Vol. 4/2, pp. 239-253, <https://doi.org/10.1177/2167702615591768>. [155]
- Vilagut, G. et al. (2013), "The Mental Component of the Short-Form 12 Health Survey (SF-12) as a measure of depressive disorders in the general population: Results with three alternative scoring methods", *Value in Health*, Vol. 16/4, pp. 564-573, <https://doi.org/10.1016/j.jval.2013.01.006>. [91]

- Vinko, M. et al. (2022), "Positive mental health in Slovenia before and during the COVID-19 pandemic", *Frontiers in Public Health*, Vol. 10, p. 3719, [103]
<https://doi.org/10.3389/fpubh.2022.963545>.
- Vistisen, H. et al. (2022), "The less depressive state of Denmark following the second wave of the COVID-19 pandemic", *Acta Neuropsychiatrica*, Vol. 34/3, pp. 163-166, [101]
<https://doi.org/10.1017/NEU.2022.1>.
- Walsh, D. and J. Foster (2021), "A call to action: A critical review of mental health related anti-stigma campaigns", *Frontiers in Public Health*, Vol. 8, p. 990, [94]
<https://doi.org/10.3389/FPUBH.2020.569539/BIBTEX>.
- Ware, J. et al. (2002), *How to score SF-12 items: How to score version 2 of the SF-12 Health Survey*, <https://www.researchgate.net/publication/291994160>. [90]
- Warr, D., R. Mann and T. Tacticos (2011), "Using peer-interviewing methods to explore place-based disadvantage: Dissolving the distance between suits and civilians", *International Journal of Social Research Methodology*, Vol. 14/5, pp. 337-352, [126]
<https://doi.org/10.1080/13645579.2010.537527>.
- Warwick Medical School (2021), *The Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS)*, <https://warwick.ac.uk/fac/sci/med/research/platform/wemwbs/>. [78]
- Woodhead, C. et al. (2012), "Impact of exposure to combat during deployment to Iraq and Afghanistan on mental health by gender", *Psychological Medicine*, Vol. 42/9, pp. 1985-1996, [106]
<https://doi.org/10.1017/S003329171100290X>.
- Yamazaki, S., S. Fukuhara and J. Green (2005), "Usefulness of five-item and three-item Mental Health Inventories to screen for depressive symptoms in the general population of Japan", *Health and Quality of Life Outcomes*, Vol. 3/1, pp. 1-7, <https://doi.org/10.1186/1477-7525-3-48>. [39]

Notes

¹ Some screening tools contain questions relating to experienced symptoms. However, we differentiate the category of “screening tools” from that of “experienced symptoms” in that the former are set piece instruments, validated against clinical diagnoses for mental health conditions, while the former are general, non-standardised question formulations asking respondents whether, for example, they “currently experience symptoms of PTSD” or “suffer from chronic anxiety” (see Table 2.7). Refer to Chapter 2 for an extended discussion on different instrument types.

² This chapter covers only four composite scales of positive mental health: the SF-12, the WHO-5, MHC-SF and (S)WEMWBS. Measurement guidelines for life evaluation, affect and eudaimonic aspects of positive mental health and subjective well-being are covered in depth in (OECD, 2013_[11]).

³ By construction, screening tools with fewer items will have lower values of Cronbach’s alpha. (Recall that Cronbach’s alpha is a function of, among other things, the total number of items in a scale.) This again underscores the importance of weighting all facets of statistical quality together, rather than placing high importance on any single test.

⁴ The two other anxiety scales against which the GAD-7 and GAD-2 were tested for convergent validity were the Beck Anxiety Inventory (BAI) and the anxiety subscale of the Symptom Checklist-90 (SCL-90).

⁵ Rasch analysis uses psychometric models to analyse categorical data and identify and measure latent attitudes or characteristics.

⁶ Although stigma and low levels of mental health literacy are strong drivers of non-response rates for mental health survey items, other factors – such as low levels of institutional trust, lack of motivation or sufficient time to participate, language barriers, poor health – may also contribute to low response rates (Lowthian and Lloyd, 2020_[162]).

⁷ Strong confidentiality assurances can reduce non-response rates for sensitive subjects (Singer, Von Thurn and Miller, 1995_[58]); however, they can in fact *increase* non-response rates for non-sensitive topics, as respondents are primed to then expect threatening or sensitive questions following an in-depth data confidentiality explanation and can be put off the interview (Singer, Hippler and Schwarz, 1992_[156]).

⁸ The correlation between risk for depressive disorders and stigma as measured through anti-stigma indicators – the share who agree that seeking treatment for mental disorders is a sign of strength, and the share who agree that mental illness is an illness like any other – show the reverse relationship, with prevalence lower in places with less bias. However, these correlations are not significant.

⁹ Cross-country and cross-group comparability are not trivial measurement issues, and some previous OECD work has dealt with the challenge of cross-country comparisons by assigning the bottom quintile of the population as at risk for mental distress, based on evidence from epidemiological studies stating that 20% of the population experiences some form of mental disorder in a given 12-month period (OECD, 2021_[57]). However, this approach by definition assigns constant prevalence, which especially in the aftermath of the COVID-19 pandemic – which saw governments across the OECD struggling to deal with huge spikes in population mental distress, depression, anxiety and stress – is limiting.

¹⁰ The geographic range where the WHO-5 has been used in surveys encompasses: Africa (Algeria, South Africa), Asia (Bangladesh, China, India, Japan, South Korea, Sri Lanka, Taiwan, Thailand), Europe (Northern, Southern, Eastern, Western and Central Europe), the Americas (Canada, the United States, Brazil, Mexico), the Middle East (Israel, Iran, Lebanon) and Oceania (Australia, New Zealand) (Topp et al., 2015_[49]).

¹¹ These emerging methods rely heavily on the application of modern psychometric methods, such as item response theory (IRT), to improve the validity, accuracy, comparability and efficiency of mental health scales, which have in turn shown substantial promise in the advanced analysis of cross-cultural differences. Using IRT-based differential item functioning as well as the use of item anchoring or equating, new methods are able to adjust for any significant bias (Dere et al., 2015_[149]; Gibbons and Skevington, 2018_[141]; Vaughn-Coaxum, Mair and Weisz, 2016_[155]). Similarly, new IRT models have emerged that can estimate and correct for extreme response styles more effectively than classical methods and quantify the tendency of extreme responding on a particular scale (Dowling et al., 2016_[150]; Jin and Wang, 2014_[152]). Some of these new methods include item banking, adapting testing and data-driven short scales and scale equating.

¹² While most international research has confirmed this rising trend of mental ill-health (OECD, 2021_[98]; Santomauro et al., 2021_[161]), evidence from individual countries at times show slightly different trajectories of mental health outcomes. A German study found that the prevalence of depression fell in the first year of the pandemic, but began rising by October 2020 and subsequently increased further over the course of 2021 and 2022 (Hapke et al., 2022_[158]; Mauz et al., 2022_[159]). An epidemiological study in Norway found that the prevalence of mental disorders decreased slightly in the early days of the pandemic (May 2020), before returning to pre-pandemic levels by September 2020 – suggesting relatively stable levels of mental disorders (Knudsen et al., 2021_[160]). This mirrors findings from a meta-analysis of 65 studies from early 2020 which showed only a small average increase in mental health symptoms in March and April 2020 that had abated by July. Both studies concluded by early Q3 2020, leaving open the possibility that an extension of the research might unveil findings similar to that of Germany – little to no change in the early days of the pandemic, but rises in distress by late 2020 and 2021.

¹³ These patterns also exist for physical health outcomes. A joint United States and Canada study found that self-administered respondents were more likely to report lower health-related quality-of-life (HRQoL) outcomes than did interviewer-administered telephone survey respondents (Hanmer, Hays and Fryback, 2007_[142]); another study in Spain found that respondents reported better physical health outcomes, measured by the SF-36, when surveys were administered by interviewers (García et al., 2005_[143]).

¹⁴ CCHS surveys include both computer-assisted personal interviews (CAPI) and computer-assisted telephone interviews (CATI). Between 2001 and 2003, the survey changed the ratio of CATI to CAPI interviews, which allowed researchers to study how mode effects affected the comparability of CCHS data across rounds. They found differences in health indicator outcomes by mode: those interviewed in person reported higher obesity rates and were more likely to be inactive, to smoke and to report contacts with medical doctors. However, self-reported mental health showed no mode effects (St-Pierre and Béland, 2004_[117]).

¹⁵ In 2020 and 2021, many countries that still use face-to-face data collection switched to telephone surveys due to the COVID-19 pandemic. These mode shifts have not been included in the figure, as mental health outcomes in these years would be heavily influenced by the global pandemic.

¹⁶ However, it is impossible to disentangle mode effects from the socio-political events that may have necessitated Gallup to change modes in the first place, which would be expected to exhibit an influence on underlying mental health. Many of the mode switches highlighted in this figure take place in countries that experienced significant political disruptions, or incidents of violence, that likely informed Gallup's choice to change the mode of data collection in the first place. For example, mode switches in Türkiye coincide with the 2016 attempted coup; the mode switch in 2013 in Iraq coincides with a ramping up of ISIS activity in the region; the mode switch in Libya coincides with the start of the second civil war; and so on. All of these events have a real impact on population negative affect balance and would very likely drive some of the changes shown in the figure, independently of mode effects.

¹⁷ For example, while mental disorders are still largely perceived as shameful in Mexico, the Mexican National Comorbidity Survey interviewers experienced few refusal rates and over the course of speaking with respondents found that people were willing to open up about their mental health, often for the first time ever (Medina-Mora et al., 2008_[153]).

¹⁸ Floor effects occur when there is bunching at the lower end of the scale, whereas ceiling effects occur when there is bunching at the upper end of the scale. In Figure 3.9, the GHQ-12 shows floor effects in that most respondents fall at the lower end of the scale, which indicates they are not at any significant risk for mental distress. Because the scale focuses on those experiencing distress, it may then be less sensitive at distinguishing between individuals with higher levels of underlying positive mental health.

¹⁹ However, some studies suggest a ceiling effect is present for (S)WEMWBS.

²⁰ Despite the rigor of the clinical validation process, criticisms of threshold scores remain. The cut-off scores that optimise sensitivity and specificity can differ – at times considerably – across population groups, and as a result alternatives to the use of cut-offs have been proposed (Goldberg, Oldehinkel and Ormel, 1998_[151]). One such proposal is the application of stratum-specific likelihood ratios, rather than fixed thresholds, so as to allow for more detailed classification systems (Furukawa et al., 2001_[144]; Furukawa and Goldberg, 1999_[145]). Additionally, new findings from research into self-reported symptoms have found that the use of single sum-scores and clinical cut-offs to estimate risk for major depression may conceal important clinical insights into depression research (Fried, 2017_[146]). To overcome these issues, some researchers have recommended the use of multiple depression scales to generate robust and generalisable conclusions, or the use of scales that include important non-DSM symptoms (e.g. the Symptoms of Depression Questionnaire (Pedrelli et al., 2013_[147])). While it is useful to note these nuances, for a government agency measuring population mental health at a macro level – as opposed to a healthcare professional at a clinical level – there is little to suggest that the use of threshold scores is inappropriate or uninformative.

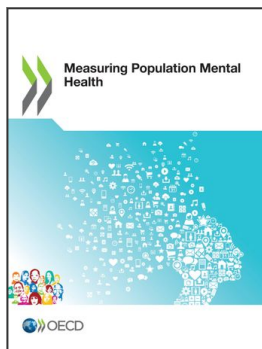
²¹ Other studies have used cut-off scores ranging from 54 to 76 (Thorsen et al., 2013_[130]; Hoeymans et al., 2004_[154]).

²² A respondent is deemed to be at risk for major depressive disorder if they answer “more than half the days” to either of the first two questions on the PHQ-8, and, in addition, a total of five or more of the eight items are reported as “more than half the days” (Eurostat, n.d._[74]).

²³ Lack of consistency in cut-off score usage can lead to confusion. For example, in 2019 three entities in Los Angeles provided wildly different estimates for the prevalence of mental health conditions among the homeless population. The *Los Angeles Times*, the Los Angeles Homeless Services Authority and the California Policy Lab at the University of California Los Angeles made estimates of 67%, 29% and 78%, respectively. All these came from the same dataset, with differences stemming from statistical interpretation (Smith and Oreskes, 2019_[157]).

²⁴ Other researchers have suggested fixed cut-off points on the SWEMWBS scale: low mental well-being (having a score between 7.00 and 19.98), moderate (19.99 to 29.30) and high (29.31 to 35.00). These categories are derived from previous work on the Danish population, with low mental well-being corresponding to the bottom 15th percentile of the distribution, and high mental well-being the top 15th percentile (Santini et al., 2022_[148]). Fixed cut-off scores have not been developed for the full-length WEMWBS.

²⁵ Care should be taken when comparing statistics on risk for major depressive disorder, or risk for depressive symptoms, coming from the PHQ-4 vs. PHQ-8 or -9. There are a number of scoring conventions for the PHQ that can lead to different prevalence estimates. Directly comparable estimates can be created by calculating risk for depression from the two individual indicators that appear in both the PHQ-4 and the PHQ-8. In this way, measures between general social and health surveys can be fully aligned, even if other (historical) health reporting has used the full set of PHQ-8 indicators to estimate depression risk prevalence.



From:
Measuring Population Mental Health

Access the complete publication at:
<https://doi.org/10.1787/5171eef8-en>

Please cite this chapter as:

OECD (2023), “Good practices for measuring population mental health in household surveys”, in *Measuring Population Mental Health*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/7c58dda7-en>

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at <http://www.oecd.org/termsandconditions>.