Common Measures in Mental Health Science:
Patient Health Questionnaire (PHQ-9)

Background

In June 2020, the National Institute of Mental Health and Wellcome reached a landmark agreement to require the use of a common set of measurement tools in the research they fund. Since then, the International Alliance of Mental Health Research Funders has supported a wide group of funders and journals to join the effort, supported by an expert advisory board.

The fundamental mission of the Common Measures in Mental Health Science initiative is to ensure that research leads to tangible improvements in the lives of people who experience mental health issues. Given the current, fragmented landscape of mental health data, there is a need to take pragmatic action to make mental health research easier to compare, communicate and interpret.

In the interest of transparency and collaboration, we are sharing the anonymised notes from the discussions that the Common Measures in Mental Health Science Advisory Committee (CMA) had on the PHQ-9 on August 25 and September 21, 2021

The CMA were asked to share:

a) Their views on the concepts that underlie the questions in the PHQ-9

b) Their views on how well these concepts are captured by the questions

c) Their experience of using these questions with different populations

d) What they thought might be missing and needed to get core concepts

Contents

The overall performance of the PHQ-9 across contexts

The PHQ was not developed as a psychometric instrument but more as a checklist for DSM.

Regarding the three different components to depression (ACE model – activity, cognition, emotion), how many are core and how many are peripheral depends on the type of depression you're trying to measure.

PHQ displays uni-dimensionality (https://psycnet.apa.org/doi/10.1037/pas0001124) and aligns with other measures (https://psycnet.apa.org/doi/10.1037/a0035768) and with DSM criteria (https://doi.org/10.1016/j.genhosppsych.2014.09.009). Cut-offs are quite consistent across different settings.

PHQ worked better than expected when used in a rural setting in Ethiopia, but one CMA member reported they had to separate items which involved multiple clauses, simplify responses, and item 8 never works very well. The validated cut-offs in rural/non-literate populations are much lower (5 or more for moderate MDD in Ethiopia and Ghana – separate validation studies).

The PHQ-9 does not have good validity or case identification when used as individual items.

It is often required (by ethics or services) to drop item 9 (suicide question) in certain settings (without referral pathways in place). It is very culturally sensitive as suicide is illegal in some religions/countries. If you ask it, you're moving from asking about mental health to religion, belief, and afterlife which can have an impact on rapport/relationship.

The PHQ-9 maps well onto the DSM construct of depression, but how valid is that construct in different cultural contexts or even within the same context?

The PHQ-9 worked well in a clinical outpatient setting in Pakistan but the main aim was that it could be used in primary care settings, which are more rural in Pakistan. It doesn't work there because of barriers to care, describing feeling 'low', and seeking out psychiatric help.

The PHQ-9 fails if we try to disaggregate it into individual items because it was intended to meet multiple DSM criteria. If individual questions were asked in other kinds of configurative groupings with other domains of life, juxtaposed with questions about sleep or eating, they would likely mean something very different.

The *Depressed* project looks at screening tools and how they perform compared with diagnostic tests. They advise against using PHQ-9 scores as an indication of diagnostic status. They have studied how PHQ items 1 and 2 perform in different contexts (https://www.depressd.ca/publishedmanuscripts)

The double-barrelled nature of many of the questions is a problem. Many scale developers avoid this fastidiously (e.g., https://doi.org/10.1016/j.jclinepi.2015.03.022, https://doi.org/10.1177%2F1073191111411667). The PROMIS depression items are simple and short.

The scale does not have an option between 'not at all' and 'several days'. There is no option for people who experience something every day.

In some contexts, it is culturally inappropriate to express depressive symptoms (e.g., studying post-natal depression in Pakistan) so have to use more somatic questionnaires like the SRQ – but PHQ is useful for cost-effectiveness evaluations, case versus non-case, and clinical thresholds.

Views on the concepts that underlie the PHQ-9 questions

Response scale:
0=not at all, 1=several days, 2=more than half the days, 3=nearly every day, 9=missing

1. Anhedonia

"Over the last two weeks, how often have you been bothered by having little interest or pleasure in doing things?"

Views on the concepts that underlie the question

The question is trying to get at anhedonia or a lack of drive to do things.

Distress caused by symptoms (and/or impact on functioning), rather than just having them.

This question, and indeed the whole of the PHQ, is an operationalisation of the DSM criteria. There was an aspiration that you could get diagnoses by giving priority to the first two core symptoms but core symptoms may not be the same across settings.

Views on how well these concepts are captured by the question

Not clear if "bothered" relates to the extent or impact of symptoms.

Is it 'level of subjective distress'? Being bothered is not a great indicator of functional impact.

Adds complexity conflating symptoms with functioning - if people have the symptom but no functional impact, how do they respond?

'Bothered by' phrasing doesn't get at whether this is a change or a more constant state. Would need a question added (for all the pool of items potentially) that asks how long these symptoms have been ongoing to get at whether this is a new incident or more of a prevalent, longer case. To address this, some CMA members reported asking about 'change' in symptoms rather than general 'bothered by'.

Experience of using this question with different populations

Many people think about this question in terms of physical health, not necessarily mental distress: 'I don't have interest' is often interpreted as 'I don't feel well'.

Regarding "pleasure" - When used in the Congo with pregnant women, they said "I understand what this means but it is just not how we think about it.

Young adults find it quite confusing to try and disentangle the symptoms from the distress caused by symptoms.

In Ethiopian studies, the data suggest that respondents just ignore the 'bothered by' bit. Tenth item on functioning is pretty useful to operationalise a 'case' rather than only the core symptoms.

What may be missing or needed to get core concepts?

The concept of rumination, or thinking too much, is missing. It is consistently found across cultures and contexts, reported by clinicians, but not part of the DSM.

There has been some in-depth work in Ethiopia and irritability comes up again and again. In international diagnostic criteria, irritability is used in depression in adolescents and older people, but not working-age people.

Noise intolerance (comes up in Ethiopia) is part of the conceptualisation of depression but not part of the DSM.

Lack/loss of motivation is one very often heard from young people.

Other missing constructs that were deemed important by CMA members: withdrawal, social isolation, anger, and irritability.

3. Sleep

Over the last 2 weeks, how often have you been bothered by trouble falling or staying asleep or sleeping too much?

Views on the concepts that underlie the question

No comments

Views on how well these concepts are captured by the question

Not sleeping enough/trouble falling asleep is driven by different factors than sleeping too much (rumination vs exhaustion). May be related to many things (see above).

Experience of using this question with different populations

No comments

What may be missing or needed to get core concepts?

We frequently simplify to 'trouble falling asleep'.

Probe further – related to generalized distress/mental health problems/stress etc. so probe what is causing the change.

Fitbits etc. not useful because the question is about 'being bothered by', not experiencing poor sleep per se. Subjective reports are typically more important for psychopathology (e.g., https://doi.org/10.5665/sleep.3834, https://doi.org/10.1016/j.smrv.2014.03.006).

Actual interaction with a phone/watch during sleep hours might be more informative.

5. Appetite

Over the last 2 weeks, how often have you been bothered by poor appetite or overeating?

Views on the concepts that underlie the question

Best understood as a measure of 'general wellness'.

Views on how well these concepts are captured by the question

Less difficulty with this than with sleep. Easier for people to understand what we're asking. But harder for people to sense their perception of appetite – particularly cross-culturally (e.g. rural Africa) therefore harder to answer.

Experience of using this question with different populations

Indigenous Australians didn't resonate with binge eating; 'sometimes it's appropriate to binge eat'.

Does the 'bothered by' get round this in different cultures? (potentially only when extreme, still harder to conceptualise).

Somatic symptoms combined (sleep and appetite)

Views on the concepts that underlie the questions

More than one somatic symptom is often indicative of an emotional problem but when taken one at a time, they become more problematic to interpret.

Concepts themselves are problematic as it is known these are key indicators of many mental health conditions and stress. They are items that show up on PTSD, anxiety, parenting measures etc. They impact so many parts of life and their specificity to depression is low. OK in a general measure but can't be pulled out.

Views on how well these concepts are captured by the questions

The somatic questions are related to nutritional deprivation and to physical health. These may have nothing to do with distress and may be due to poor nutrition, lack of food, physical health ailments, HIV, and other illnesses.

Experience of using these questions with different populations

In perinatal women in Ethiopia objectively defined measured anemia was unrelated to fatigue but depressive symptoms were more strongly predictive.

9. Suicide and suicidal ideation

"Over the past 14 days, how often have you been bothered by thoughts you would be better off dead or of hurting yourself in some way?"

Note: Often excluded in population settings where no resource to offer support to people who endorse this. Also cultural issues.

Views on the concepts that underlie the question

Gut feeling is it should always be used. Safety plans if interview in person. Not hard to develop.

Also other approaches. Can send out pre-emptive plans with resources but make it clear there isn't a way of responding in the moment of answering the question.

Ethical implications of asking, but also ethical implications of not asking. People think it might increase suicidal thoughts by being asked, but evidence (https://doi.org/10.1016/j.cpr.2018.07.001, https://doi.org/10.1016/j.psychres.2017.08.048, https://doi.org/10.1080/13811118.2020.1793857) has found the opposite (more common to report less distress (relief) than more after being asked)

Views on how well these concepts are captured by the question

Conflates suicidal ideation with self-harm ideation which can be quite different. Doesn't assess duration or intensity.

Experience of using this question with different populations

Big cultural issues e.g. risk of involuntary psychiatric holds etc. (and religion/legality points mentioned above) may lead to non-response.

Adaptations reported by CMA members

The coupling of overeating with undereating, as well as oversleeping with undersleeping is problematic in Ethiopia because of the connotations of oversleeping with laziness and overeating with greed. Particularly a problem if 'negative' extreme is second, and if aural rather than reading on the page as the second example stands out more. This makes the respondent less likely to endorse the statement, so CMA member reports splitting them. They report that whilst this causes some difficulties in the field for scoring, it does work (scale formatted in this way demonstrated to be valid in Ethiopia and Ghana).

By adding the functioning criteria in the form of the tenth item of the PHQ-9, some CMA members reported that they find that is a better way of operationalising a case rather than running the algorithm and prioritizing the core symptoms.

Can break scale responses into yes/no, then (if yes) level, in the context of older adults in rural Indonesia, and in Ethiopia and Ghana.

It is critical to identify pattern variations in context, such as a group responding under famine conditions vs high-income middle class. Is that contextual data being collected, mandatory, and what data do we need to correctly interpret variations?

Not clear that specifying time period matters. One CMA member reported that when comparing scales head-to-head, some using present state, some asking about symptoms over the last month, they all perform the same against a gold standard criterion of MDD. One CMA member reported from one of their studies "when we asked about the 2-week scale and how people answered it - several said that it was mostly driven by how they were feeling on the day of the interview/assessment". They noted that in their experience, people tend to rapidly answer the items and do not do much interpretation. They make a rough estimate of the days. This may be why different scales work similarly.

Advice for the direction of the Common Measures initiative

If it is decided that the PHQ-9 isn't a good instrument going forward… we should not feel like we have to throw away legacy data. Some items might become less useful for comparisons. Could also create cross-walking studies to relate new data from different measures to those collected with legacy tools.

PHQ-9 can already be scored on the PROMIS metric, which defines depression more precisely with emotional and cognitive symptoms (but doesn't have sleep). This would allow us to work between measures, including legacy data (but then have to do x-cultural work with PROMIS).

Cross-walking has been done, enabling us to relate new data from different measures to data collected with legacy tools: e.g. https://doi.org/10.1037/a0035768.

Changing items in the scale may require revalidating the whole thing, which may not be in the remit of the initiative. Some disagreement that splitting an item may not invalidate the scale.

Some combined items are problematic in different cultures, splitting these items can improve appropriateness, potentially without requiring full revalidation.

Splitting questions may be fine in terms of psychometrics (not aware of this having been empirically tested) - but when wording is changed (even slightly) we often see major differences in item responses, discrimination, etc.

Cultural differences are a problem when comparing scores between different cultures but aren't much of a problem within a population if everyone makes the same error. There seems to be very little work on the impact of statistical differences in scales across cultures. Is it a problem or not? (Depressed project found that doing fancy latent modelling with weights didn't make a difference)

It is critical to identify pattern variations in context, such as a group responding under famine conditions vs high-income middle class. Is that contextual data being collected, mandatory and what data do we need to correctly interpret variations? Also relevant to functioning – measures need to map together to provide the relevant context.

References

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, PROMIS Cooperative Group. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment. 2011 Sep;18(3):263-83.

Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychological assessment. 2014 Jun;26(2):513.

Maglione JE, Ancoli-Israel S, Peters KW, Paudel ML, Yaffe K, Ensrud KE, Stone KL. Subjective and objective sleep disturbance and longitudinal risk of depression in a cohort of older women. Sleep. 2014 Jul 1;37(7):1-9.

Lovato N, Gradisar M. A meta-analysis and model of the relationship between sleep and depression in adolescents: recommendations for future research and clinical practice. Sleep medicine reviews. 2014 Dec 1;18(6):521-9.

Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. General hospital psychiatry. 2015 Jan 1;37(1):67-75.

Batterham PJ, Brewer JL, Tjhin A, Sunderland M, Carragher N, Calear AL. Systematic item selection process applied to developing item pools for assessing multiple mental health problems. Journal of clinical epidemiology. 2015 Aug 1;68(8):913-9.

Hanlon C, Medhin G, Selamu M, Breuer E, Worku B, Hailemariam M, Lund C, Prince M, Fekadu A. Validity of brief screening questionnaires to detect depression in primary care in Ethiopia. Journal of affective disorders. 2015 Nov 1;186:32-9.

Smartt C, Medhin G, Alem A, Patel V, Dewey M, Prince M, Hanlon C. Fatigue as a manifestation of psychosocial distress in a low-income country: a population-based panel study. Tropical Medicine & International Health. 2016 Mar;21(3):365-72.
Batterham PJ, Calear AL, Carragher N, Sunderland M. Prevalence and predictors of distress associated with completion of an online survey assessing mental health and suicidality in the community. Psychiatry research. 2018 Apr 1;262:348-50.

Blades CA, Stritzke WG, Page AC, Brown JD. The benefits and risks of asking research participants about suicide: A meta-analysis of the impact of exposure to suicide-related content. Clinical psychology review. 2018 Aug 1;64:1-2.

Wu Y, Levis B, Riehm KE, Saadat N, Levis AW, Azar M, Rice DB, Boruff J, Cuijpers P, Gilbody S, Ioannidis JP. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychological medicine. 2020 Jun;50(8):1368-80.

Tekola B, Mayston R, Eshetu T, Birhane R, Milkias B, Hanlon C, Fekadu A. Understandings of depression among community members and primary healthcare attendees in rural Ethiopia: a qualitative study. Transcultural Psychiatry. 2020:13634615211064367.

Bianchi R, Verkuilen J, Toker S, Schonfeld IS, Gerber M, Brähler E, Kroenke K. Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. Psychological Assessment. 2022 Jun;34(6):595.

Habtamu K, Birhane R, Medhin G, Hanlon C, Fekadu A. Psychometric properties of screening questionnaires to detect depression in primary healthcare setting in rural Ethiopia. BMC Primary Care. 2022 Jun 2;23(1):138.

Polihronis C, Cloutier P, Kaur J, Skinner R, Cappelli M. What's the harm in asking? A systematic review and meta-analysis on the risks of asking about suicide-related behaviors and self-harm with quality appraisal. Archives of Suicide Research. 2022 Apr 3;26(2):325-47.

Fekadu A, Demissie M, Birhane R, Medhin G, Bitew T, Hailemariam M, Minaye A, Habtamu K, Milkias B, Petersen I, Patel V. Under detection of depression in primary care settings in low and middle-income countries: a systematic review and meta-analysis. Systematic Reviews. 2022 Dec;11(1):1-0.