An anthropological critique of psychiatric rating scales

ARTICLE

Neil Armstrong & Nicola Byrom

SUMMARY

This article discusses sceptical arguments about measurement scales. Measurement scales are part of a promising agenda of openness, transparency and patient and public involvement (PPI) in medical research, but have received critical, sometimes hostile attention from anthropologists. This is because scales repackage localised cultural assumptions about distress as something universal and pan-human and have the capacity to reshape people's interior lives in unhelpful, possibly harmful ways. We take as an example the Patient Health Questionnaire-9 (PHQ-9). Use of the PHQ-9 is currently mandated by major funders. But its history suggests flawed PPI and a lack of openness. The article suggests a constructive role for anthropology in mental health research, using ethnographic evidence and theory to show how, although they have their uses, mental health scales should not be regarded as inert or harmless.

LEARNING OBJECTIVES

After reading this article you will be able to:

- understand the use, and some anthropological critiques, of measurement scales
- outline the history of the PHQ-9
- appreciate patients' experiences of measurement scales within mental healthcare.

KEYWORDS

Anthropology; rating scales; depressive disorders; history of psychiatry; patients and service users.

To understand the significance of the measurement of distress and well-being, we need to understand the context. Medical and psychological research today values three principles in particular: patient and public involvement (PPI), open science and the standardisation of measurement tools. The PPI agenda is refreshing. Working with, and listening to, patients and the public in general appears likely to enhance the reliability, validity, relevance and usefulness of research. For this reason, funders such as Wellcome now require PPI in their mental health research stream (Wellcome 2023). The open science agenda recognises and appreciates transparency. This means research protocols should be shared in advance and data should be made openly accessible. Finally, standardisation allows better comparison between studies, facilitating meta-analysis, which allows researchers to see the bigger picture and uncovers biases. For example, the Wellcome Trust now expects all researchers studying depression to use the Patient Health Questionnaire-9 (PHQ-9), as an agreed measure of depression (Wellcome 2023). If we all use the same measure, we all measure the same thing, allowing maximum comparability of data. So, the use of explicit, standardised measurement scales to quantify distress and well-being is integral to the contemporary research process.

At an abstract level, each of these research principles has merit. They might seem hard to disagree with. Who could declare themselves in favour of excluding the public from scientific research, actively concealing data or promoting the use of idiosyncratic, single-use outcome measures? Scientific research is ultimately collaborative, and all three principles – PPI, transparency and standardisation – promote cooperation that is likely to lead to improved research. In mental healthcare research, we might expect this to lead to higher quality care, more effective interventions and improved outcomes. But when researchers start to enact the principles, the situation becomes murkier. How should they be put into practice? Are they mutually consistent? It is not clear that scales like the PHQ-9 that purport to measure distress do so in an objective, causally neutral and culture-free way. Indeed, thinking of distress as a personal, subjective, measurable experience is contentious. Raising such concerns need not be seen as an attack. Rather, thinking them through might be a means of enriching or enhancing current practice.

Anthropology and the measurement of distress

The idea that fundamental human experiences, such as love, desire or distress, might be variable across populations is nothing new. Charles Darwin noted cross-cultural variation in human emotions over time (Darwin 1872). Subsequent historical and conceptual work has suggested that our current notions of symptoms are 'unstable constructs' (Marková 2009). Anthropological research offers additional

Neil Armstrong, MA, MSt, DPhil, is a lecturer in anthropology at SOAS University of London (School of Oriental and African Studies), London, UK. **Nicola Byrom**, BSc, DPhil, is a senior lecturer in the Department of Psychology, King's College London, London, UK. **Correspondence** Neil Armstrong. Email: na72@soas.ac.uk

First received 16 Oct 2023
Final revision 25 Jul 2024
Accepted 10 Sep 2024

https://doi.org/10.1192/bja.2024.60 Published online by Cambridge University Press

ways of addressing these issues both empirically and conceptually. Anthropologists have used cross-cultural ethnographic evidence to investigate measurement scales, both as technologies that generate knowledge and as instruments in use in day-to-day mental healthcare. In a number of studies from the 1980s onwards, anthropologists showed how apparently straightforward scientific terms such as 'mood' and 'depression', critical components of psychiatric scales, are not universal or pan-human, but are cultural and localised. Arthur Kleinman and Byron Good, for example, demonstrated that both the way depression was understood and how it was experienced differed in significant ways in different societies (Kleinman 1985). In a study of emotional life on a Pacific island, Catherine Lutz found people's interior worlds so unfamiliar to her as a North American that she was able to conclude: 'emotional experience is not precultural, but *preeminently* cultural' (Lutz 1988: p. 5). Intimate, interior experiences such as sadness or depression or despair may be subjective and involuntary, but they do not arise directly out of our human nature and they are felt differently according to the social context.

This has been seen as a problem for measurement scales in psychiatry, because they adopt a single way of expressing affective states and then assume they hold for all individuals in all cultures. For the PHQ-9 (Box 1) 'Trouble concentrating on things, such as reading the newspaper or watching television' and 'Feeling bad about yourself – or that you are a failure or have let yourself or your family down' can be indicative of a depressive disorder. The idea that all humans are able to experience depressive disorders and that these disorders involve trouble concentrating on things and feeling bad about yourself is far from certain. In anthropological terms, measurement scales 'naturalise' a particular set of locally contingent affective values and practices, making them seem inevitable and obvious and unquestionable, almost as if scales are not made by researchers but exist independently and are discovered by them.

An earlier generation of anthropologists regarded the discovery of cross-cultural variation in our emotional lives as intrinsically hostile to the psychological and psychiatric sciences. It was seen as an epistemic scandal that mental health professionals were foisting cultural artefacts such as scales on unsuspecting patients. Drawing on the work of Foucault, mental healthcare became understood as a form of power, a way of containing and disciplining unruly citizens through dominating knowledge (Rose 1999). This kind of argument puts anthropology into an antagonistic relationship with psychiatry and psychology. However, this hostility is not inevitable.

BOX 1 The Patient Health Questionnaire-9 (PHQ-9)

The first part of the PHQ-9 (each item scored 0–3: Not at all/Several days/More than half the days/Nearly every day) asks:

'Over the last 2 weeks, how often have you been bothered by any of the following problems?:

1 Little interest or pleasure in doing things
2 Feeling down, depressed, or hopeless
3 Trouble falling or staying asleep, or sleeping too much
4 Feeling tired or having little energy
5 Poor appetite or overeating
6 Feeling bad about yourself – or that you are a failure or have let yourself or your family down
7 Trouble concentrating on things, such as reading the newspaper or watching television
8 Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual
9 Thoughts that you would be better off dead or of hurting yourself in some way'

The second part (using a tick-box response: Not difficult at all/Somewhat difficult/Very difficult/Extremely difficult) asks:

'If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?'

The full PHQ-9 is freely available at www.phqscreeners.com/select-screener.

More recently, anthropologists have begun to explore other lines of enquiry. It is not that the cultural particularity or causal efficacy of measurement scales has come under question. Rather, anthropologists have begun to think that even if the biomedical sciences underestimate both the influence of culture and the malleability of humans, this need not discredit the whole of mental healthcare. Anthropologists such as Joanna Cook and Liana Chase suggest a more complex causal world in which the cultural effects of mental healthcare might be positive as well as negative, helping people to thrive rather than imposing the will of the state (Chase 2021; Cook 2023). Emily Martin, an anthropologist who has been diagnosed with bipolar disorder, notes positives and negatives in the way scales become part of a person's life-world: 'The individual uniqueness of experience might be lost in the homogenising process of abstraction, but in return, private moods take the form of their opposite, moods that are widely shared' (Martin 2007: p. 195). For Martin, scales

BJPsych Advances (2024), page 1 of 9 doi: 10.1192/bja.2024.60

decontextualise distress, stripping it of meaning and unjustly increasing a sense of personal responsibility by erasing structural disadvantage, but they also make painful interior experiences articulable, connect people and promote solidarity. These more recent publications point to the possibility of a more productive relationship between anthropology and psychiatry. We return to these issues shortly, but first explore the background to one scale in particular, the PHQ-9.

The social and cultural origins of the PHQ-9

The PHQ-9 is a curious measure because although its use supports the principle of standardisation, it is unclear how well it meets the needs of the open science and PPI agenda. We focus on it here not to suggest that scales in general (or the PHQ-9 in particular) are wrong, or invalid or without value. Rather we want to demonstrate by means of a single example how scales have a history and a social life that should not be ignored or erased.

Constructing a scale

Scales are typically used to represent a behaviour, a feeling or an action that cannot be captured in a single variable or item (Boateng 2018). To construct a scale, researchers progress through a series of phases, including item development, scale development and scale evaluation. Item development should include inductive methods, using exploratory research methods including focus groups and interviews. With items developed, the validity of the content must be assessed to test the adequacy with which the measure assesses the domain of interest (Hinkin 1995). This is deemed vital if the items are to measure what they are presumed to measure. This phase should include both evaluation by experts who are highly knowledgeable about the domain of interest and evaluation by the target population. Boateng et al (2018) describe an iterative process in which rounds of Delphi consultation with experts are interspersed with focus groups with the target population to work towards a consensus on the definition of the domains being measured and the possible items to include. For a measure of depression this would mean repeated consultation with research experts and psychiatrists, interspersed with careful consultation with individuals with lived experience of depression, to identify a shared understanding of depression and the facets that it feels helpful to measure within a scale. The subsequent phase of scale development should include pretesting the questions through cognitive interviews with the target population prior to administering the survey and working through phases of item reduction, factor extraction and tests of reliability and validity (Kimberlin 2008; Boateng 2018). In addition to content validity, the criterion and construct validity of a scale must be considered (Kimberlin 2008; Boateng 2018). Criterion validity can be taken as the extent to which a measure predicts a result with which it ought to be related (predictive validity) and the extent to which test scores relate to established measures of the same construct (concurrent validity; Boateng 2018). Construct validity is the extent to which an instrument assesses a construct of concern (Raykov 2010).

The PRIME-MD: the precursor of the PHQ-9

The PHQ-9 as we know it today was derived from the PRIME-MD (Primary Care Evaluation of Mental Disorders), an interview prompt for physicians unsure of how to assess, within the time constraints of a busy clinic, whether a patient was experiencing a psychiatric condition (Spitzer 1994, 1999). A group of psychiatrists, funded by the pharmaceutical company Pfizer, determined a list of questions designed to identify a range of mental disorders, including depression, anxiety and eating disorders, somatoform disorders and probable alcohol misuse. Notably, the original publication paper is not an open access manuscript and therefore cannot legally be read without payment by, for instance, most patients being screened for depression. The study focuses on assessing whether the PRIME-MD can be administered more quickly than previous approaches, whether it helps primary care physicians reach the same conclusions as mental health practitioners, increasing the recognition of mental disorders by primary care physicians and whether patients believe the tool helps their physician understand their problems. None of the information reported in the paper bears any resemblance to the process currently accepted as required to develop a new measurement scale (see also Morgado et al, 2017). Instead, the items were developed from the DSM-III-R (American Psychiatric Association 1987). In the introduction, the paper notes an 8-month development phase, including 450 patients, in which interviews were conducted to revise the items. However, no details of this phase are described in the paper, let alone any data from this process being made openly available.

Predictive, construct, content and criterion validity of the PHQ-9

Thousands of papers cite the PHQ-9 and many of these consider the predictive validity of the scale, assessing the ability of the scale to arrive at the same diagnostic conclusion as that made through an in-depth interview (e.g. Wu 2020; Costantini

2021). At least in high-income Western countries there is a considerable weight of evidence to show that the PHQ-9 usually produces the same diagnostic conclusion as an extended interview (e.g. Wu 2020; Costantini 2021; Negeri 2021). Notably, these evaluations are narrow in focus, in comparison, for instance, with a comprehensive review of the reliability and validity of the Hamilton Rating Scale for Depression (Bagby 2004).

There are relatively few qualitative evaluations of the patient experience of completing this scale. These do suggest reasonable construct validity; looking at patterns of change in global PHQ-9 scores, Malpass et al (2010) found convergence with patient descriptions of their experience of symptoms. However, the scale may not accurately capture suicidality since suicidal ideation and self-harm are underreported on the scale (Malpass 2010; Richards 2019).

Qualitative evaluations suggest some limitations of content validity. Discussing the original development of the PHQ-9, the team note that 89% of patients believed that the questions were 'very' or 'somewhat' helpful in getting their physicians to better understand or treat the problems they were having (Spitzer 1999). This points to a form of content validity. Subsequent qualitative studies have found patients noting that many symptoms and experiences that are meaningful to them are not captured on the scale, including libido, sense of vacancy, sense of time, irritation, fluctuations in mood, awareness and sense of stability (Malpass 2010). Others have reflected that the scale fails to capture how patients truly feel (Richards 2019) or that patients find themselves unable to 'fit' their experience into the response options and therefore often feel the questionnaire is misrepresenting their experience (Malpass 2016).

There are strong correlations between scales commonly used to measure depression. This would often suffice for confirmation of criterion validity. However, when considering individual cases there are discrepancies such that 'scales are "reading" the depressive features in these cases differently and to an extent that would alter clinical interpretation' (Hawley 2013). These data casts wider doubt on the overall validity of depression rating scales as idiographic measures of depression, prompting the caution that a scale, such as the PHQ-9, should be used in conjunction with 'an experienced psychiatrist's global assessment of depression severity' (Hawley 2013). A similar recommendation for a more holistic analysis was made by Costantini et al (2021); following their systematic review of the use of the PHQ-9 in primary care, they recommend coordination between mental health and primary care services to support positive PHQ-9 screening results to be rapidly followed up with a structured diagnostic interview.

The purpose of the PHQ-9

The purpose for which a scale is developed is important. Hamilton (1976) set this out at the genesis of modern scale development: 'there are four types of scales in psychiatry: for assessment of the patient's condition, for diagnosis, for prognosis and for the selection of treatment'. Hamilton suggests that scales should not be used interchangeably: scales that are good for one purpose may not be suitable for other purposes. The PHQ-9 was developed expressly for the purpose of diagnosing mental disorders in primary care (Spitzer 1999).

The original publication of the PHQ-9 reports on the successes of this scale in terms of time savings (Spitzer 1999). This is a very genuine concern: Spitzer and colleagues noted that time limitations in a busy office setting may be a major obstacle to the recognition of mental disorders. Lack of resourcing for primary care remains a major barrier to identification of mental health problems, especially in low- and middle-income countries (Knapp 2006). Time constraints are frequently noted as a significant diagnostic challenge in primary care settings (e.g. Campbell 2001; Wilson 2002; McKenna 2004; Fleury 2012; Kroenke 2017; Rogers 2021). Spitzer and colleagues describe how the PHQ-9 can be completed in around 1 min, a major reduction from the 8 min it took doctors to work through the PRIME-MD. The focus for development of the PHQ-9 was on the utility of the scale for doctors. The authors acknowledge that there is a downside to this more time-efficient process: as the PHQ-9 is a self-administered scale, the doctor's engagement with the patient is reduced, leading to a less detailed or comprehensive understanding of the patient, and a relationship with diminished therapeutic potential (Spitzer 1999). Subsequent qualitative research with practitioners echoes the concern of intruding on the delivery of individualised patient care (Leydon 2011). In practice, clinicians in general practice tend to adapt the PHQ-9 for use in interviews with patients, at times making strategic use of the scale to overcome resistance to forms of treatment (Ford 2020).

Functional impairment: an overlooked indicator in the PHQ-9

The original PHQ asked respondents about the impact of problems on day-to-day life: 'How difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?' This item alone provided a good global prediction of the likelihood of a clinical diagnosis (Spitzer 1999). This item behaves very differently from the other PHQ items. While the PHQ items try to pin down and define the shape of a

BJPsych Advances (2024), page 1 of 9 doi: 10.1192/bja.2024.60

problem, this item allows the problem to be undefined, and asks simply about the impact or consequence of this undefined distress. Sadly, this item, although valuable in predicting diagnosis, does not feature in routine use of the PHQ-9 today. As clinicians and researchers seek to define the shape of problems, it is important not to lose sight of what really matters to individuals, the impact of these 'problems' on their lives. Reflecting on the rising rates of mental health problems among young people, Jack Andrews and Susanne Schweizer stress the importance of taking such a functional measurement, recommending that functional impairment, the extent to which an individual's symptoms affect their daily functioning, should be the focus of mental health interventions for young people (Andrews 2023).

What is it like to be defined by the PHQ-9?

'Looping' and 'discursive operators'

Researchers from the humanities and social sciences have investigated the effect of questionnaires on a person's life. In a series of influential publications in the 1990s, philosopher Ian Hacking suggested that medical terms can 'loop' back and change the people they set out to describe (Hacking 1995, 1998). This means calling an unhappy person 'depressed' might change the nature of their sadness. Equally, declaring that someone is 'psychotic' will affect the person in question rather differently from describing them as being 'possessed by spirits'. For Hacking this is not trivial. He uses historical material to show that certain forms of distress, such as dissociative fugue states and multiple personality, appear to come and go, as if by contagion. Fugue states became common in late 19th-century Western Europe, but then more or less died out. In a similar way, in the late 20th-century USA, diagnosis of multiple personality appears to have clustered around clinics that treated multiple personality. The prevalence of these forms of distress appears to rest on many factors, which, when they co-occur, form what Hacking calls an 'ecological niche'. One of the factors is looping. What Hacking has in mind is that certain ways of thinking about, for example, multiple personality contribute to experiences of multiple personality. So how we represent distress in medical notes, in diagnostic manuals and in symptom scales is not clinically irrelevant. It can become entwined in the phenomenology of distress.

Hacking's work is hugely valuable, but might feel a little distanced from lived experience. Anthropologists Janis Jenkins and Thomas Csordas have tried to capture the personal impact of medical terminology in terms of 'discursive operators' that become part of the fabric of lived experience: 'Regardless of whether the discourse of diagnosis is relevant in any given moment of experience or social interaction, it is most certainly sedimented into the lifeworld [of patients]' (Jenkins 2020: p. 128). Jenkins & Csordas argue that medical apparatus such as scales are themselves efficacious. They change a person. Rebecca Lester makes a similar point about how eating disorders are not wholly separable from treatments for eating disorders but meet and merge and become folded together in complex and sometimes hard-to-predict ways (Lester 2019). It remains an open question, of course, as to whether sedimenting clinical categories into the life-world of patients is, on balance, beneficial or harmful. Psychologists such as Lucy Foulkes have suggested that efforts to raise awareness of mental health problems may have significantly contributed to the rise of mental health problems in the Western world (Foulkes 2023).

'Value capture' and the harmful potential of metrics

One reason why researchers became concerned about the causal properties of metrics is that they are seen as necessarily simplistic, a means of reducing a unique and complex human life to a single, manageable number. This often means simplifying nuanced and heterogeneous experiences such as depression (Goldberg 2011; Rantala 2018), giving them the appearance of a unitary disorder. The philosopher C. Thi Nguyen has recently tried to explore the harm this might do by developing the concept of 'value capture' (Nguyen 2024). Nguyen suggests that if a person starts out wanting to exercise because it is relaxing, improves health, enhances sleep, offers a chance to reflect and promotes a sense of connection with nature, but ends up just chasing 10 000 daily steps on their phone's step counter, they may be victims of value capture. Value capture occurs when a person exchanges their own nuanced and evolving values (in this case regarding exercise) and replaces them with something simplistic, inflexible and prefabricated (such as a numerical target). Hitting or not hitting a step target is a very poor proxy for the range of benefits a person might get from exercise. Late-night running up and down stairs might enable a person to meet their step count, but it may not help them reflect on their life and certainly will not promote a sense of connection with nature. If a person wants to focus on the full range of benefits from exercise, they might do well to either ignore their step count altogether or at least not allow it to erase other values. Nguyen notes that external bodies (such as insurance companies, which work

at the level of populations, not individuals) might like metrics, and they have become essential to the workings of contemporary capitalism. But none of that is a good reason for an individual to adopt them in place of their own rich, subtle, context-specific values. When they ask patients to compete the PHQ-9, clinicians may be engaging in value capture. A person's own sense of their life, their trajectory and commitments and emotions are first reduced to nine questions and then to a single number.

The patient's experience of scales: the PHQ-9

Both of us have experience of completing measures like the PHQ-9 as mental healthcare patients. We describe these experiences in Boxes 2 and 3. Neither of us found it comfortable. We were not told that the scale was designed to speed things up in the clinic and to provide support to physicians who felt they needed assistance. The above-mentioned publications by Hacking, Jenkins & Csordas, Lester and Nguyen suggest some provocative questions. We might recognise that something is lost if a person who exercises because it is relaxing, has health benefits, enhances sleep, offers a chance to reflect and promotes a sense of connection with nature ends up just chasing 10 000 daily steps on a step counter. But what might be lost if a person starts to take the PHQ-9 to be more authoritative than, or preferable to, their own intuitions? What happens if discursive operators from the PHQ-9 become sedimented into a person's life-world or if the PHQ-9 loops back to reshape the people it describes?

The PHQ-9 only allows for reporting of distress in a way that is decontextualised. It starts out by describing feeling tired, or feeling down or feeling bad about yourself as 'a problem' which a person may have been 'bothered' by. It asks in general how often the respondent has felt down, depressed or hopeless. Negative emotional states are problematic when they are felt to be personal or internal, rather than circumstantial or external, permanent rather than temporary, and pervasive rather than

BOX 2 Lessons patients learn from psychiatric scales: an autoethnographic account by Neil Armstrong

When I was diagnosed with bipolar disorder, I wasn't just offered treatment or support. I was offered an education. I soon learned that alongside clinical skills, mental health professionals are teachers, and that measurement scales like the PHQ-9 form part of the curriculum. I was taught that my inner world contained moods that go up and down. This wasn't immediately obvious to me. I saw it more as different kinds of subjective state that shifted and changed over time. But mood scales don't accommodate ideas like that. They work by extracting numerical answers to questions, such that inner states can be represented in terms of quantity. So, I was taught that moods fluctuate like oil prices, and that when they get too high or too low, they can interfere with life and thus become symptoms of a mental disorder. It suggested to me a previously unsuspected kinship between my mental distress and world economic crises.

I was amazed at the reductive power of scales. Instruments to measure symptoms taught me that a wide range of experiences should be framed as instances of depression: poor concentration, guilt, weight gain, restlessness, fearfulness, directionlessness and angst. To me, each was distinct, encompassing not just feelings felt in the moment, but beliefs, ethical commitments and tastes. But the PHQ-9 thought otherwise. Even concern about the impact of my disorder on my future career registered on a mood scale and thus became folded into the disorder itself.

Over the course of my patient life, I completed scales many times. The sheer frequency of them was irksome. Measurement made care seem crude, mechanical and anything but person-centred. At best, they were a mandatory intrusion on clinical relationships that might be aimed at the therapeutic. In response I became disingenuous and strategic. It was extremely frustrating because I didn't really agree with the assumptions that lay behind the scales. I didn't believe that my interior life could be measured like this, and, when presented with the results of the scale I found it unconvincing, outlandish, even bizarre. It didn't matter. I was told that everyone has mental health and that moods and anxiety going up and down are a fixed and universal part of our evolved human nature. Poor concentration, restlessness and angst may feel different, but this is just surface gloss. They are, underneath, and at heart, naturally part of a broader umbrella concept of depression. It was demonstrated by science, they said. If I expressed sceptical views, I was turned into a contrarian, or a dissident. I knew I had to be careful so as not be labelled by clinicians as non-compliant. Yet, in the research literature, the implicit message of scales was neither obvious nor uncontested. In fact, my scepticism was widely shared by researchers. So the curriculum I was taught by scales and measures is just one way among many of thinking about distress.

Continual exposure to scales like the PHQ-9 impaired my capacity for reflection. The consequences of this are something I describe in a recent book: 'Once diagnosed, it became axiomatic for me that there was little or no relationship between the external world and my internal world. My bipolar self was socially differentiated, sealed off and self-generated, discontinuous with those round me' (Armstrong 2023: p. 102). These effects turned out to be harmful because they undermined my confidence and formed a barrier to personal growth or maturation and, thus, to recovery. In order to get well, I needed to find alternate ways to think of my distress. I need to own it, to see it as arising out of my life and my values. This demanded something much deeper than mood scores or outcomes measures.

BOX 3 What the PHQ-9 misses in depression: an autoethnographic account by Nicola Byrom

My experience of mental illness has never been tidy. I've had periods of my life where I have not been very happy and did not like myself much. Bluntly, I thought the world would be better of if I was dead. I have had persistent thoughts of hurting myself. As with many individuals struggling with their mental health, admitting that I was struggling, let alone contemplating that I might benefit from, or even need, help, was very hard. I have desperately needed someone to stop me and say 'I do not think you are okay'. While I've seen enough clinicians, and indeed seen clinicians at times when they were checking in with me to establish if I was doing okay, it has never been a clinician who has said 'I do not think you are okay'.

A clinician's primary tool to identify whether I needed help was the PHQ-9, or often the two-item screening, PHQ-2. The phrasing of these scales perplexes me: according to these scales I may not have been depressed. I might have done nothing kind for myself and not allowed myself time to do anything enjoyable, but these negative experiences would usually have been balanced out by the apparent 'positives' of my extreme interest in my work or trying to care for a new baby. At periods of acute distress, I could have told you that I spent most of the day crying, but I'm not sure I felt down, depressed or hopeless. I felt sad, unsettled, unable to cope and lost. Had a clinician moved on to the PHQ-9, I have had times where all I wanted to do was sleep: it provided escape from daily life. But I would have been the worst judge of whetr that was too much sleep. I might have been very tired and had little energy, but admitting this to myself, let alone another person, would have been a profound admission of failure. And talking about failure, asking about whether you feel like a failure out of context is flawed. Take time to talk to me about how I was doing, and you would have learnt that I felt like an acute failure in some areas of my life, so I pushed myself to extremes to perform in other areas to mitigate this failure. It was not healthy. It was not sane. But it would not show up on the PHQ-9.

Completing the PHQ-9 at periods of my life when I have felt most distressed would have failed to help a clinician identify my depression. Worse, as an accessible tool, it has been something to check to allow me to tell myself that, whatever chaos might be going on in my mind, I am not depressed and so do not need help. The voices of mental illness are cruel and uncaring, they are entirely devoid of compassion. Emotional state in illness can be labile and chaotic, confusing and complex. I've never felt it to fit neatly into the PHQ's definitions of depression. It never felt that straightforward. In this sense, attempts to distil assessment of mental distress into nine or even two simple questions have done me immeasurable harm, both failing to see the crisis in its midst and helping me deny that crisis.

specific. The questions strip away meaning from distress, rendering painful inner states mysterious and baffling, as if they were unwanted aliens that come and go according to their own agenda. Depression is made boundless and patients become strangers to themselves, deskilled, disempowered, even helpless (Rotter 1966; Abramson 1978; Maier 2016). There is no value placed on curiosity about the background causes or the nature of painful emotions. What matters is the quantity.

So, if we were on the lookout for looping effects, or discursive operators or value capture, we might be looking for people who, after being subject to the PHQ-9 for a significant length of time, are inclined to see strong feelings as something of a bother, a health problem that might interfere with their life, almost external to them, somehow not arising out of their values and experiences. Readers can judge for themselves whether the two autoethnographic accounts contain evidence of these effects.

Conclusion

The current research agenda in psychiatry sounds convincing, but anthropologists have raised concerns about how it plays out in practice. There is necessarily a degree of speculation here, not because such effects are inaccessible to research, but because existing research efforts have been directed elsewhere. But there are plenty of reasons for concern. Even a cursory look at the development of the PHQ-9 reveals internal problems. By mandating use of the PHQ-9, funders are able to standardise research. However, since the scale was created for the purpose of maximising time efficiency for unsure physicians and developed with vague PPI and very limited transparency, it creates contradictions between the principles that research funders are advocating. We might want to put it more bluntly: to insist on the use of the PHQ-9 while also advocating for PPI feels like gaslighting patients. It suggests that patient involvement is important in shaping research, but professionals should be left with the task of quantifying the shape of the patient's distress. When we were asked to complete a PHQ-9, the scale was presented as being for the benefit of patients, not clinicians. Imposing the scale had relational and communicative consequences. The PHQ-9 might help a clinician to defend a decision and to produce high-quality record of their care, but it does so in a way that is consequential and, at very least, requires analysis.

More broadly, two key arguments made by anthropologists suggest an urgent need for further research. First, the anthropological work by Kleinman, Good and Lutz reviewed above showed that feelings are culturally contingent. This indicates

MCQ answers
1 d 2 b 3 e 4 c 5 c

that scales and other paraphernalia of psychiatric life reflect a particular cultural take on distress, something not universal but localised and contingent. Other ways of thinking about distress, other cultures of distress, are not only possible, but actually found in cross-cultural data. Psychology research moves in a similar direction when it suggests that feeling down or hopelessness can be meaningful signals, communicating that there are problems with facets of the individual's life that need to be addressed (Gut 1989; Nesse 2000; Forgas 2017). Feeling hopeless when thinking about one's career might be a meaningful call to action, to explore career change, for example. Feeling unconcerned about climate change or the military situation in Ukraine might indicate a loss of compassion that hints at a deeper despair.

Second, cultures of distress are causally active. They change people and their experiences. As a component of our biomedical culture of distress, the PHQ-9 has an impact on the life-world of patients. It loops back to affect what it feels like to be unhappy. It might, even, replace nuanced, flexible, personal understandings of distress with simplistic, perhaps crude questions and categories. If the PHQ-9 is an instance of value capture, we might expect it to harm patients. Knowledge oriented towards clinical utility for psychiatrists might simultaneously also be disempowering, even unhealthy, for patients. At present, mental distress and disability due to mental distress are rising. Interdisciplinary research is needed to investigate to what degree measurement scales are part of the cure, or part of the problem.

Author contributions

This article was commissioned by *BJPsych Advances* and authorship is shared equally between N.A. and N.B.

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Declaration of interest

None.

References

Abramson LY, Seligman ME, Teasdale JD (1978) Learned helplessness in humans: critique and reformulation. *Journal of Abnormal Psychology*, **87**: 49–74.

American Psychiatric Association (1987) *Diagnostic and Statistical Manual of Mental Disorders* (3rd edn revised) (DSM-III-R). APA.

Andrews JL, Schweizer S (2023) The need for functional assessments in school-based mental health intervention research. *JAMA Psychiatry*, **80**: 103–4.

Armstrong N (2023) *Collaborative Ethnographic Working in Mental Health*. Routledge.

Bagby RM, Ryder AG, Schuller DR, et al (2004) The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *American Journal of Psychiatry*, **161**: 2163–77.

Boateng GO, Neilands TB, Frongillo EA, et al (2018) Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health*, **6**: 149.

Campbell SM, Hann M, Hacker J, et al (2001) Identifying predictors of high quality care in English general practice: observational study. *BMJ*, **323**: 784–7.

Chase L (2021) Psychosocialization in Nepal: notes on translation from the frontlines of global mental health. *Medicine Anthropology Theory*, **8**: 1–29.

Cook J (2023) *Making a Mindful Nation: Mental Health and Governance in the Twenty-First Century*. Princeton University Press.

Costantini L, Pasquarella C, Odone A, et al (2021) Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): a systematic review. *Journal of Affective Disorders*, **279**: 473–83.

Darwin C (1872) *The Expression of the Emotions in Man and Animals*. John Murray.

Fleury MJ, Imboua A, Aubé D, et al (2012) General practitioners' management of mental disorders: a rewarding practice with considerable obstacles. *BMC Family Practice*, **13**: 19.

Ford J, Thomas F, Byng R, et al (2020) Use of the Patient Health Questionnaire (PHQ-9) in practice: interactions between patients and physicians. *Qualitative Health Research*, **30**: 2146–59.

Forgas JP (2017) Can sadness be good for you? *Australian Psychologist*, **52**: 3–13.

Foulkes L, Andrews JL (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*, **69**: 101010.

Goldberg D (2011) The heterogeneity of "major depression". *World Psychiatry*, **10**: 226–8.

Gut E (1989) *Productive and Unproductive Depression: Success or Failure of a Vital Process*. Basic Books.

Hacking I (1995) *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton University Press.

Hacking I (1998) *Mad Travellers: Reflections on the Reality of Transient Mental Illness*. University of Virginia Press.

Hamilton M (1976) Comparative value of rating scales. *British Journal of Clinical Pharmacology*, **3**(suppl 1): 58–60.

Hawley CJ, Gale TM, Smith PS, et al (2013) Equations for converting scores between depression scales (MÅDRS, SRS, PHQ-9 and BDI-II): good statistical, but weak idiographic, validity. *Hum Psychopharmacol*, **28**: 544–51.

Hinkin TR (1995) A review of scale development practices in the study of organizations. *Journal of Management*, **21**: 967–88.

Jenkins J, Csordas T (2020) *Troubled in the Land of Enchantment: Adolescent Experience of Psychiatric Treatment*. University of California.

Kimberlin CL, Winterstein AG (2008) Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, **65**: 2276–84.

Kleinman A, Good B (1985) *Culture and Depression: Studies in the Anthropology and Cross-Cultural Psychiatry of Affect and Disorder*. Berkely.

Knapp M, Funk M, Curran C, et al (2006) Economic barriers to better mental health practice and policy. *Health Policy and Planning*, **21**: 157–70.

Kroenke K, Unutzer J (2017) Closing the false divide: sustainable approaches to integrating mental health services into primary care. *Journal General Internal Medicine*, **32**: 404–10.

Lester R (2019) *Famished: Eating Disorders and Failed Care in America*. University of California Press.

Leydon GM, Dowrick CF, McBride AS, et al (2011) Questionnaire severity measures for depression: a threat to the doctor–patient relationship? *British Journal of General Practice*, **61**: 117–23.

Lutz C (1988) *Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory*. University of Chicago.

Maier SF, Seligman ME (2016) Learned helplessness at fifty: insights from neuroscience. *Psychological Review*, **123**: 349–67.

Malpass A, Shaw A, Kessler D, et al (2010) Concordance between PHQ-9 scores and patients' experiences of depression: a mixed methods study. *British Journal of General Practice*, **60**: e231–8.

Malpass A, Dowrick C, Gilbody S, et al (2016) Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. *British Journal of General Practice*, **66**: e78–84.

Marková IS, Berrios GE (2009) Epistemology of mental symptoms. *Psychopathology*, **42**: 343–9.

Martin E (2007) *Bipolar Expeditions: Mania and Depression in American Culture*. Princeton University Press.

McKenna HP, Ashton S, Keeney S (2004) Barriers to evidence-based practice in primary care. *Journal of Advanced Nursing*, **45**: 178–89.

Morgado FF, Meireles JF, Neves CM, et al (2017) Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, **30**(1): 3.

Negeri ZF, Levis B, Sun Y, et al (2021) Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. *BMJ*, **375**: n2183.

Nesse RM (2000) Is depression an adaptation? *Arch Gen Psychiatry*, **57**: 14–20.

Nguyen CT (2024) Value capture. *Journal of Ethics and Social Philosophy*, **27**: 469–504.

Rantala MJ, Luoto S, Krams I, et al (2018) Depression subtyping based on evolutionary psychiatry: proximate mechanisms and ultimate functions. *Brain, Behavior, and Immunity*, **69**: 603–17.

Raykov T, Marcoulides GA (2010) *Introduction to Psychometric Theory*. Routledge.

Richards JE, Hohl SD, Whiteside U, et al (2019) If you listen, I will talk: the experience of being asked about suicidality during routine primary care. *Journal of General Internal Medicine*, **34**: 2075–82.

Rogers R, Hartigan SE, Sanders CE (2021) Identifying mental disorders in primary care: diagnostic accuracy of the Connected Mind Fast Check (CMFC) electronic screen. *Journal of Clinical Psychology in Medical Settings*, **28**: 882–96.

Rose N (1999) *Governing the Soul: The Shaping of the Private Self* (2nd edn). Free Association Books.

Rotter JB (1966) Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, **80**(1): 1–28.

Spitzer RL, Williams JBW, Kroenke K, et al (1994) Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. *JAMA*, **272**: 1749–56.

Spitzer RL, Kroenke K, Williams JB, et al (1999) Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA*, **282**: 1737–44.

Wellcome (2023) *Common Metrics in Mental Health Research*. Wellcome (https://wellcome.org/grant-funding/guidance/common-metrics-mental-health-research).

Wilson A, Childs S (2002) The relationship between consultation length, process and outcomes in general practice: a systematic review. *British Journal of General Practice*, **52**: 1012–20.

Wu Y, Levis B, Riehm KE, et al (2020) Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychological Medicine*, **50**: 1368–80.

MCQs

Select the single best option for each question stem

1 Cross-cultural research by anthropologists has found that:
- a all cultures experience depression in the same way
- b hatred is only felt by humans in the northern hemisphere
- c emotions can be felt by cats
- d there is significant cross-cultural variation in emotions
- e depression is the same the world over.

2 The PHQ-9 was developed to:
- a enhance accuracy in diagnosis
- b save time in the clinic
- c make diagnosis more reliable
- d improve clinical relationships
- e help insurance companies with their paperwork.

3 The PHQ-9 includes questions on:
- a feeling morally compromised by the demands of work or home life
- b having a sense that life is meaningless
- c hearing voices
- d lacking a sense of community
- e thinking that you might be better off dead.

4 The autoethnographic accounts in this article of filling in scales like the PHQ-9 suggest that:
- a scales promote better communication between patient and clinician
- b decontextualising distress as a personal, internal symptom promotes insight
- c the kind of knowledge required by clinicians to promote evidence-based decision-making may not the same kind of knowledge required by patients to make sense of their experiences
- d scales have no impact on a patient
- e patients are unconcerned about a mismatch between complex subjective states and scale items.

5 The current research agenda in psychiatry:
- a is focused on answering fundamental questions about human existence
- b is based on a rich account of human flourishing
- c prioritises openness, transparency and patient and public involvement
- d prioritises interdisciplinarity, mystery and uncertainty
- e uses concepts derived from South Asia.