The Journal of Child Psychology and Psychiatry

Journal of Child Psychology and Psychiatry **:* (2020), pp **–** doi:10.1111/jcpp.1333

Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement

Praveetha Patalay,[1] and Eiko I. Fried[2]

[1]Centre for Longitudinal Studies and MRC Unit for Lifelong Health and Ageing, University College London, London, UK; [2]Clinical Psychology Unit, Leiden University, Leiden, The Netherlands

Introduction

In July 2020, two of the largest funders of mental health research worldwide – the National Institute of Mental Health (NIMH) and the Wellcome Trust – announced plans to standardize mental health measurement (Farber et al., 2020). Specifically, obtaining funding for research related to depression and anxiety will be conditional on using four specific measures (Farber et al., 2020; Wolpert, 2020). This is especially relevant given that Wellcome recently identified mental health as a strategic priority area and committed £200 million to depression and anxiety research in young people.

The measures being mandated are (a) Patient Health Questionnaire (PHQ-9) for depression, (b) General Anxiety Disorder (GAD-7) for anxiety, (c) Revised Child Anxiety and Depression Scale (RCADS-22) for depression and anxiety in children and adolescents, and (d) World Health Organisation Disability Assessment Schedule (WHODAS) for impact on adult functioning.

While we agree that there are obvious benefits to standardizing mental health measurement, some of which are discussed in the announcement by NIMH and Wellcome (Wolpert, 2020), in this paper we focus on potential unintended negative consequences of this initiative (also summarised in Box 1), and layout recommendations for how some of these might be mitigated (also summarised in Box 2).

A. Lacking transferability across settings

There is an abundance of measures for specific domains of mental health. For example, over 280 scales have been used to measure depression in the last century (Santor et al., 2006). While numerous scales exist for similar purposes – for example a recent systematic review of depression trials identified 19 different outcome measures across 30 trials (Mew et al., 2020) – the diversity of measures in part also reflects the different needs of research and practice. [Correction made on 30 Sep 2020, after first online publication: In the preceding sentence 'meta-analysis' has been corrected to 'systematic review']. Very brief scales can be used in the emergency room where there may only be time for one or two questions; other scales were developed to assess severity of symptoms or monitor treatment progress in already diagnosed patients; and some scales are screeners for mental health problems in general population settings.

The PHQ-9 – the measure mandated by NIMH and Wellcome for depression – falls into the third category. It was not created with the purpose to, for example, track depression severity in patients during treatment, and as such, the scale may lack important psychometric properties in clinical samples, such as unidimensionality (Titov et al., 2011) and measurement invariance (i.e. measuring the same construct in different populations) (Baas et al., 2011). As a short screener, the scale also only provides limited insight into the types and extent of patients' difficulties (Fried, 2017) and only correlates moderately with other depression scales developed for clinical settings (Wittkampf et al., 2009).

Scales are developed for certain settings or purposes. Settings include clinics, schools or the general population, and purposes encompass situations in which measures are used, such as observational research, intervention evaluation or routine patient monitoring. We note that in over 30 population-based studies in the United Kingdom, none include the RCADS or WHODAS; 2 include the GAD-7; and 2 include the PHQ-9 (Catalogue of Mental Heath, 2019). Similarly, although the focus of Wellcome's strategic priority investment is on depression and anxiety in young people, a systematic review of school-based intervention studies for depression and anxiety in young people indicates that none of 81 identified studies used the RCADS, and only one used PHQ-9 and GAD-7 (Werner-Seidler et al., 2017). This is because these scales were not developed – and have not been used widely to date – in these settings and for these purposes. Time frame of assessment is another relevant consideration. The

Conflict of interest statement: No conflicts declared.

Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA 02148, USA

PHQ-9 queries participants about symptoms in the last 2 weeks, and differs from assessments of symptoms in the last few hours for momentary assessments, or assessments of symptoms in the last year or even during lifetime for estimating population or lifetime prevalence (Box 1).

There are no objective measures of mental health; existing measures have specific properties and were designed for certain settings and purposes. There is insufficient evidence that all four prescribed scales have the sort of transferability that would make them good measures across various contexts. Scale validity, reliability and utility, as well as further considerations such as acceptability to respondents, should be demonstrated across settings and purposes before they are mandated for universal use.

B. Narrowing the scope of inquiry

Mood and anxiety disorders are highly heterogeneous, and different individuals can suffer from very different sets of symptoms (Fried & Nesse, 2015). In addition, they are often very broad constructs. For instance, common scales for measuring depression encompass over 50 disparate symptoms (Fried, 2017).

The scales mandated by NIMH and Wellcome assess nine symptoms of depression (PHQ-9), 7 symptoms of anxiety (GAD-7) and 22 symptoms across both in RCADS. Hence, these scales can only provide limited insights into the full range of difficulties individuals might experience. While this will undoubtedly lead to useful information on these specific symptoms across multiple settings, it risks sidelining all the other ways in which people experience distress. Some of the difficulties not included in these scales might be crucial targets for treatment or understanding aetiology, and standardizing measurement to brief assessments risks that widespread data collection efforts overlook these problems and risk missing important insights.

In addition to being broad and complex constructs, mental disorders are highly comorbid, and their risk factors are often transdiagnostic. This contrasts with the notion of many separate, clearly circumscribed, categorical diseases, as is portrayed in widely used diagnostic manuals such as the Diagnostic and Statistical Manual of Mental Disorders (DSM). Over the last decades, researchers have regarded the DSM with increasing scepticism, and there has been growing consensus that categorical DSM disorders and their accompanying symptoms have considerable limitations. One of the limitations of such diagnostic manuals is the narrowly defined scope of each disorder. For example, major depressive disorder entails only nine depressive symptoms – not by accident nearly identical with the symptoms in the PHQ-9 mandated by NIMH and Wellcome – and fails to capture many other problems relevant to the wider depressive syndrome, such as anxiety and anger that are highly prevalent and associated with worse clinical outcomes (Fava et al., 2008; Judd et al., 2013). The decision to mandate scales like the PHQ-9 comes at a time where the field has widely acknowledged limitations inherent to DSM's conceptualization of mental health disorders in general and major depression specifically, and we see the grave risk of rolling back years of progress and consensus building around limitations of DSM categories.

Mood and anxiety disorders constitute a wide umbrella of difficulties and are among the leading causes of disease burden worldwide. Reducing their scope to a few specific symptoms means turning a blind eye to the complexity and breadth of mental health problems, limiting important insights for research and treatments while reaffirming contested diagnostic categories that the field is ready to move beyond.

Importantly, while NIMH and Wellcome did not mandate that *only* these scales be used and encourage use of additional scales alongside these scales, this does not alleviate the concerns we raise. First, there are contexts in which these scales might simply not be a good choice (due to the time frame or validation population; see issue A). Second, in contexts where limited time/resources for measurement are available (such as large-scale population-based studies), or contexts where long assessments put severe burden on respondents, researchers and practitioners, adding extra scales over the mandated ones will often not

Box 1

Unintended negative consequences of mandating standardized measures:

a. *Lacking transferability across settings:* scales were developed for specific settings (e.g. community, clinic) and purposes (e.g. intervention studies), and their properties might not be easily transferable between settings.
b. *Narrowing the scope of inquiry:* individuals experience mental health difficulties in wide-ranging ways, and the narrow scope of the proposed scales risks limiting important insights for research and treatments.
c. *Lowering the threshold for robust evidence:* empirical findings limited to a specific imperfect measure are less robust than if such evidence is (re)produced across multiple scales.
d. *Creating a two-tiered mental health science:* arbitrarily conferring gold standard status on some imperfect measures over others will create an artificial two-tiered system leading to an impoverishment of mental health research.

be possible, and using scales more suited to the populations and context will be more valuable.

C. Lowering the threshold for robust evidence

Important decisions such as approving a new treatment must be based on robust evidence. Robust here means that the finding is well established, for instance because it is replicated across a number of independent clinical trials. One important pillar of robust science is measurement. Because there are no objective measures of mental health, and because each measure is imperfect, covers only a certain range of difficulties and was developed for a specific context, evidence that a treatment works can only be considered robust if the effect generalizes across several measures.

Let's take an example. Suppose two very similar studies on the efficacy of a new depression treatment in young adults use different measures and come to different conclusions on whether the treatment works or not; it is possible that the discordant findings are due to the different measures used, an unsatisfying situation that could be avoided by standardizing measurement. However, now suppose both studies used the same measure and reached the same conclusion; while their results would align, we would be unaware of the fact that the positive finding for this treatment is dependent on using a particular scale (e.g. because it happens to cover some of the symptoms the treatment works for, or because it fails to cover important problems that get worse during treatment). In such a scenario, would we really want this treatment to be rolled out in health services to all patients?

While standardizing mental health measurement to increase comparability is a laudable aim, we fear that mandating specific imperfect measures will come at the cost of magnifying scale-specific issues and limiting the robustness of findings.

D. Creating a two-tiered mental health science

Finally, the initiative of NIMH and Wellcome confers gold standard status on specific imperfect measures. One unintended consequence of this is that it may undermine the future utility of existing studies or data sets that do not include these measures. That is, the initiative risks creating a two-tiered system whereby funders are likely to favour certain studies or health systems because they already use the mandated measures, although they may not be superior in terms of scientific quality or utility. This could encourage the sidelining of excellent and necessary depression and anxiety research that goes beyond the narrow scope of these measures. In addition, researchers may be tempted to change measures to accommodate the mandate (e.g. to secure funding), with adverse outcomes. Changing measures, for example in clinical outcome monitoring or long-term population-based longitudinal cohort studies, interrupts temporal continuity and threatens scientific utility, with no advantages for patients, clinicians or wider society.

Even if NIMH and Wellcome themselves demonstrate some amount of flexibility and discernment in their decisions around how strictly they apply this (Farber et al., 2020), we see a very real risk that other gatekeepers of mental health research and treatment delivery, including governments, funding bodies, international organizations, health system providers and scientific publishers, will not be so accommodating. Once such a mandate has become widely recognized, institutions and grant reviewers will be more likely to treat research as being more fundable because they feature one imperfect measure over another. This could quickly spill out into impacting what journals publish, narrowing the fields of inquiry around mood and anxiety disorders to a limited set of constructs (issue B) lacking contextual transferability and content validity (issue A), while simultaneously lowering the threshold for robust evidence (issue C).

Recommendations

To mitigate the potentially negative unintended consequences of the initiative by NIMH and Wellcome, our recommendations are as follows. First, given specific measures work best in specific settings and for specific purposes, we suggest mandating a wider set of recommended measures. This will allow greater flexibility to maximize scientific utility across diverse contexts while minimizing some of the issues outlined above, such as magnifying scale-specific problems and decreasing the robustness of future evidence. Second, we recommend assessing the validity, utility and transferability of measures across settings before their use is mandated. These efforts could benefit from funding specifically allocated to measurement research, such as testing whether specific scales measure the same construct across diverse populations, their sensitivity to change in different contexts and so on. Focus should be given to the most common settings and those where there is minimal prior precedence for scientific utility and validity evidence of any prescribed measures. Third, NIMH and Wellcome should more clearly stress the limitations of mandated measures to ensure that other stakeholders and gatekeepers of mental health science do not *en masse* insist on the application of this mandate across all their studies, which would reduce the quality and robustness of future mental health research.

Overall, we greatly appreciate that NIMH and Wellcome plan to review and potentially revise their decision in the future. However, we fear that this measurement mandate will be adopted so quickly that once the ball is rolling, reversing this decision will not be easy. The DSM is a good example of how once the ball starts rolling, even with the best

Box 2

Recommendations

1. Mandate a wider set of measures that have been validated for specific populations and research purposes
2. Fund research assessing the measurement properties of scales across settings and purposes
3. Stress the limitations of mandated measures to avoid *en masse* application and replacement of measures across studies and health systems
4. Create speedbumps to ensure that any widespread adoption of mandated measures does not result in impoverishment of mental health science

intentions to keep re-evaluating, decisions can be difficult to reverse. Creating 'speed bumps' in the roll-out process, including time to evaluate the impacts of this decision on research and practice, may help avoid some of the consequences we highlight in this article.

We conclude that while motivated by the right concerns around the use of multiple measures in current mental health research, the unintended consequences of mandating imperfect measures risk leading to a mental health research field that becomes conceptually poorer and analytically less robust in the coming decade. Actions to mitigate these are necessary and urgent.

Acknowledgements

No specific funding was sought for this work. P.P. and E.F. conceived the paper and were both equally responsible for writing, revising and approving the final version for submission. The authors would like to thank Drs. Suzanne H Gage (University of Liverpool) and Jan R Boehnke (University of Dundee) for their helpful feedback on an earlier draft. The authors have declared that they have no competing or potential conflicts of interest.

Correspondence

Praveetha Patalay, Centre for Longitudinal Studies and MRC Unit for Lifelong Health and Ageing, University College London, Gower Street, London WC1E 6BT, UK; Email: p.patalay@ucl.ac.uk

Key points

- We fear that mandating a limited set of imperfect mental health measures that lack contextual transferability and content validity will have multiple adverse consequences, including magnifying scale-specific issues, reaffirming contested diagnostic hegemonies and creating a two-tiered system of research and evidence in mental health science.
- If not mitigated, this will lead to narrowing the fields of inquiry around mood and anxiety disorders and lowering the threshold for robust evidence, which are highly undesirable outcomes for individuals suffering from mental health difficulties.
- Action needs to be taken urgently to mitigate these consequences and should include mandating a wider set of measures, establishing the measurement properties of prescribed scales across various settings and creating mechanisms to prevent the impact of impoverishing mental health science through a narrow set of imperfect measures.

References

Baas, K.D., Cramer, A.O., Koeter, M.W., van de Lisdonk, E.H., van Weert, H.C., & Schene, A.H. (2011). Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). *Journal of Affective Disorders*, *129*, 229–235.

Catalogue of Mental Health (2019). Catalogue of mental health measures. ESRC and CLOSER. Available from https://www.cataloguementalhealth.ac.uk/

Farber, G., Wolpert, M., & Kemmer, D. (2020). *Common measures for mental health science laying the foundations*. Wellcome Trust. Available from https://wellcome.ac.uk/sites/default/files/CMB-and-CMA-July-2020-pdf.pdf

Fava, M., Rush, A.J., Alpert, J.E., Balasubramani, G.K., Wisniewski, S.R., Carmin, C.N., … & Trivedi, M.H. (2008). Difference in treatment outcome in outpatients with anxious versus nonanxious depression: A STAR* D report. *American Journal of Psychiatry*, *165*, 342–351.

Fried, E.I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197.

Fried, E.I., & Nesse, R.M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR* D study. *Journal of Affective Disorders*, *172*, 96–102.

Judd, L.L., Schettler, P.J., Coryell, W., Akiskal, H.S., & Fiedorowicz, J.G. (2013). Overt irritability/anger in unipolar major depressive episodes: past and current characteristics and implications for long-term course. *JAMA Psychiatry*, *70*, 1171–1180.

Mew, E.J., Monsour, A., Saeed, L., Santos, L., Patel, S., Courtney, D.B., … & Butcher, N.J. (2020). Systematic scoping review identifies heterogeneity in outcomes measured in adolescent depression clinical trials. *Journal of Clinical Epidemiology*, *126*, 71–79.

Santor, D.A., Gregus, M., & Welch, A. (2006). Eight decades of measurement in depression. *Measurement: Interdisciplinary Research and Perspectives*, *4*, 135–155.

Titov, N., Dear, B.F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cognitive Behaviour Therapy*, *40*, 126–136.

Werner-Seidler, A., Perry, Y., Calear, A.L., Newby, J.M., & Christensen, H. (2017). School-based depression and anxiety prevention programs for young people: A systematic review and meta-analysis. *Clinical Psychology Review*, *51*, 30–47.

Wittkampf, K., van Ravesteijn, H., Baas, K., van de Hoogen, H., Schene, A., Bindels, P., . . . & van Weert, H. (2009). The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *General Hospital Psychiatry*, *31*, 451–459.

Wolpert, M. (2020). Funders agree first common metrics for mental health science. Available from https://www.linkedin.com/pulse/funders-agree-first-common-metrics-mental-health-science-wolpert

Accepted for publication: 21 August 2020