PERSONAL VIEW

Have our Cake and Eat it too? The Challenge of Ensuring Affordability, Sustainability, Consistency, and Adaptability in the Common Metrics Agenda

Karolin Rose Krause, Sophie Chung, Maria da Luz Sousa Fialho, Peter Szatmari, Miranda Wolpert

Author Affiliations

Cundill Centre for Child and Youth Depression, Centre for Addiction and Mental Health (CAMH), Toronto, ON, Canada (KR Krause PhD, Prof P Szatmari MD).

Research Department for Clinical, Educational and Health Psychology, University College London, London, United Kingdom (KR Krause; Prof M Wolpert DClinPsy)

Independent Researcher, London, United Kingdom (S Chung BA)

International Consortium for Health Outcomes Measurement, London, United Kingdom (ML Sousa Fialho DPhil)

Department of Psychiatry, Hospital for Sick Children and the University of Toronto, ON, Canada (P Szatmari)

Wellcome Trust, London, United Kingdom (M Wolpert)

Corresponding Author: Karolin R Krause, Cundill Centre for Child and Youth Depression, Centre for Addiction and Mental Health, 80 Workman Way, Toronto, ON M6J 1H4, Canada, karolin.krause@camh.ca.

Accepted for publication by the Lancet Psychiatry on 18/03/2021.

Author Contributions

KRK, SC, MLSF, and MW conceived of the presented ideas. KRK, MLSF, and MW conducted relevant background research. KRK drafted the manuscript with technical and intellectual input from all authors. All authors contributed to refining of the manuscript. All authors have read and approved the final version of the manuscript.

Declaration of interests

KRK reports personal fees from the International Consortium for Health Outcomes Measurement (ICHOM), outside the submitted work: From October 2018 through March 2020, KRK was a research fellow with ICHOM and supported the development of a Core Outcome Set for children and youth experiencing anxiety and/or depression. KRK received personal fees from ICHOM during this period. KRK is involved with the International Network for Research Outcomes in Adolescent Depression Studies (IN-ROADS) initiative at the Hospital for Sick Children (Toronto, ON) that aims to develop a Core Outcome Set specifically for youth depression clinical trials. SC was an employee of ICHOM until December 2019 and responsible for managing the development of Core Outcome Sets in mental health. MLSF is employed by ICHOM and oversees the development of Core Outcome Sets in mental health. PS has received royalties from Guilford Press for his book *A Mind Apart: Understanding Children with Autism and Asperger Syndrome*. PS is a co-investigator for the IN-ROADS initiative that aims to develop a Core Outcome Set specifically for youth depression clinical trials. MW is Director of Mental Health at the Wellcome Trust and as such has issued recommendations for common metrics in anxiety and depression research. MW chaired (but did not vote with) the before-mentioned ICHOM working group that developed core metrics for anxiety and depression in children and young people. MW was formerly Director of the Child Outcomes Research Consortium that made recommendations about outcome measures. MW was involved in developing one of the measures recommended by the ICHOM working group—the Current View tool. This tool is free to use, and MW makes no financial gain from its use.

Acknowledgments: We thank the reviewers of this manuscript for their valuable feedback and suggestions.

Funding

The authors did not receive any grant or funding that was specifically dedicated to the preparation of this manuscript. KRK is funded by the Cundill Centre for Child and Youth Depression at the Centre for Addiction and Mental Health, in Toronto (ON, Canada). PS has received grant funding support from the Canadian Institutes of Health Research (CIHR) and the Jamie and Patsy Anderson Chair in Child and Youth Mental Health.

Key Words: Common metrics, anxiety, depression, outcome measurement, standardization, harmonization, item response theory

Word Count: 4,498 words.

Abstract

Mental health research grapples with research waste and stunting of field progression caused by varied and inconsistent outcome measurement across studies and clinical settings which means there is no common language for considering findings. Whilst recognizing there are no gold standard measures and all existing ones are flawed in one way or another, anxiety and depression research is spearheading a common metrics movement to harmonize measurement, with several recent initiatives recommending the consistent use of specific scales to allow read across in terms of measurement between studies. For this approach to flourish, however, common metrics must be acceptable and adaptable to a range of contexts and populations and access should be as easy and affordable as possible globally, including in very low resource settings. Within a measurement landscape dominated by fixed proprietary measures and with competing views of what should be measured, this poses a range of challenges. In this personal view we consider tensions between affordability, sustainability, consistency, and adaptability that, if not addressed, may risk undermining the common metrics agenda. We outline a three-pronged way forward that involves funders taking more direct responsibility for measure development and dissemination; a move towards managing measure dissemination and adaptation via open access measure hubs; and transitioning from fixed questionnaires to item banks. We argue that it is time to start thinking of mental health metrics as 21st-century tools to be co-owned and co-created by the mental health community with support from dedicated infrastructure, coordinating bodies, and funders.

Introduction

The design and provision of evidence-based mental health care depends on the availability of reliable, valid, and clinically relevant outcome data—both in research studies and as part of measurement-based care (1). While much of medical research relies on established biomarkers or accepted metrics (e.g., Body Mass Index), reliable bio markers are yet to be identified for all common mental health conditions (2). In their absence, a variety of psychometric scales serve to capture symptom clusters, functional impairment, quality of life, or overall well-being.

Over 280 scales have been developed over the past century to capture depressive symptoms alone (3); a recent scoping review of youth depression trials identified 19 different instruments used to measure depression symptom severity across 30 trials (4); and another review identified 30 different instruments used to assess anxiety symptoms in children and youth across 257 clinical trials and observational studies (5). There is a high degree of inconsistency in the types of symptoms assessed by different scales (6,7). A review of 126 questionnaires used to screen for common mental health conditions showed low rates of cross-scale symptom similarity, which ranged from 29% for bipolar disorder to a maximum of 58% for obsessive-compulsive disorder (7). The psychometric properties of different instruments also vary, as do assessment timelines and the informants who are consulted (i.e., clinicians, parents, and children/youth). The result is a "Tower of Babel"-like evidence base that hampers the synthesis and comparability of research findings via meta-analyses, pooled data analysis, or the benchmarking of outcomes across services or systems. The result is research waste (2,4,8–10), and the stunting of progress in mental health research, including for common conditions like anxiety and depression (4,6,8–10).

Several recent initiatives have aimed to overcome this state of fragmentation by recommending core metrics or Core Outcome Sets (COS) that should be administered, as a minimum, across all research studies or practice settings for a given condition (11,12). The International Consortium for Health Outcomes Measurement (ICHOM) has convened several working groups to develop COS for use in mental health care settings. As of March 2021, ICHOM sets were available for anxiety and depression in children, youth, and adults (5,13), as well as psychotic disorders, personality disorders and addiction in adolescents and adults (https://www.ichom.org/standard-sets/#mental-health). Efforts to develop COS specifically for clinical trials of youth and adult depression treatments are ongoing (10,14). In addition, leading mental health funders have agreed on a set of common metrics for mental health that should be measured in all studies conducted with their support (15), and UNICEF has led an initiative to identify, adapt, and validate consensus measures for adolescent mental health for use in population surveys worldwide (16).

Scales recommended by COS and similar initiatives are typically selected based on specific feasibility and/or psychometric criteria. The exact criteria may vary depending on the intended use context (e.g., trials versus measurement-based care), but they often include a consideration of affordability. Affordability is an important factor in ensuring the widest possible uptake of common metrics, as cost is one important barrier to the implementation of measurement-based care in practice settings (17–20), and can also influence the selection of measurement scales in clinical trials or population surveys.

All authors of this commentary have been involved in one or several of the above-mentioned common metrics initiatives. Through this involvement and conversations with tool developers, we have come to realize that within a measurement landscape dominated by fixed, proprietary, and copyrighted scales, the double aim of promoting consistency and affordability brings inherent tensions related to the adaptability and sustainability of common metrics. Below we discuss how these tensions play out in current models of measure development and dissemination, and how they could ultimately undermine the goals of the common metrics agenda. We then suggest a three-pronged way forward. While these questions are of relevance to the wider field of mental health research, we will focus on anxiety and depression metrics for children and youth as a case example.

A Landscape of Fixed and Proprietary Measures

Measurement scales typically consist of a fixed set of items that can be summed to a total score (or several subscale scores), and are usually copyrighted to guarantee developers adequate monetary or non-monetary (e.g., attribution) compensation in exchange for allowing others to utilize their intellectual property (21). Scales come with varying conditions for reuse, modification, translation, and further dissemination (Table 1). Reliable costing, and licencing information is often difficult to find, and may be hidden behind the paywalls of commercialized measure catalogues. Researchers and practitioners who use or adapt a measure without adhering to licencing terms risk legal pursuits, barriers to publishing the resulting research, or calls to retract already published research (28,29).

Table 1. Levels of Cost and Control over Outcome Measures in Mental Health as of March 2021

Use incurs a fee [a]	Controls are placed on tool adaptation, translation and/or dissemination	Description of the Dissemination Model	Example
Yes	**Yes**	**Cost and control:** A limited set of manuals and questionnaires can be purchased from the copyright holder at a cost. Permission must be obtained to reproduce any copyrighted material. See: https://www.pearsonassessments.com/footer/permissions---licensing.html	Beck Youth Inventories ™ (22) (managed by Pearson)
Case by case	**Yes**	**Partially free use; limited usage:** The free use, copy and reproduction of HoNOSCA materials without express permission is allowed for care providers within the United Kingdom's National Health Service and "in other countries such as Australia, New Zealand and Switzerland, where HoNOS has been mandated for use to support assessment and outcome monitoring in public and private sector mental health services.". Otherwise, explicit permission must be obtained from the Royal College of Psychiatrists. Developer consent is needed to copy, distribute, or adapt the scale for non-commercial use: "The following acts may not be performed without the consent of the Royal College of Psychiatrists: Copying the work; Renting or otherwise issuing copies of the work to the public; Adapting the work [e.g. changes to the wording of items, scaling of items, addition or deletion of items or changes in the order of items]." See: https://www.rcpsych.ac.uk/events/in-house-training/health-of-nation-outcome-scales	HoNOSCA (23) (managed by the Royal College of Psychiatrists)
No	**No**	**Free for all users and usages:** Tool is in the public domain, and free for any use: "All PHQ, GAD-7 screeners and translations are downloadable from this website and no permission is required to reproduce, translate, display or distribute them". See: https://www.phqscreeners.com/select-screener	GAD-7 and PHQ-9 (24,25) (managed by Pfizer Inc.)

Use incurs a fee [a]	Controls are placed on tool adaptation, translation and/or dissemination	Description of the Dissemination Model	Example
No	**Yes**	**Free use, limited usage**: Free for non-commercial use but users are required to sign a user agreement. Written permission required for modification, adaptation, or translation: "User shall not modify, abridge, condense, translate, adapt, recast or transform the WHODAS 2.0 in any manner or form, including but not limited to any minor or significant change in wording or organization, or administration procedures, of the WHODAS 2.0. If User thinks that changes are necessary for its work, or if translation is necessary, User must obtain written approval from WHO in advance of making such changes." [Paragraph within the user agreement form]. See: https://www.who.int/classifications/icf/WHODAS_2_0_UserAgreement_v2011.pdf	WHODAS 2.0 (26) (managed by the WHO)
No	**Yes**	**Free use, collaborative ongoing development**: "All English and Spanish versions of PROMIS […] are publicly available for [single] use without licensing or royalty fees for individual research or individual clinical use". "User agrees not to adapt, alter, amend, abridge, modify, condense, make derivative works, or translate HealthMeasures Instruments without prior written permission from the Provider. In cases where permission is granted, User will be expected to evaluate the impact of approved modifications." "clinical researchers are encouraged to submit de-identified data for collaborative analysis and reporting. […] Clinical researchers are strongly encouraged to collaborate with HealthMeasures investigators when applying these items and banks to their research". See: http://www.healthmeasures.net/images/PROMIS/Terms_of_Use_HM_approved_1-12-17_-_Updated_Copyright_Notices.pdf	PROMIS (27) (managed by a network of primary research sites and coordinating centres)

Note. GAD-7: Generalized Anxiety Disorder 7-item Scale; HoNOSCA: Health of the Nation Outcome Scales for Children and Adolescents; PhQ-9: Patient Health Questionnaire; PROMIS: Patient-Reported Outcomes Measurement Information System; WHO: World Health Organization; WHODAS 2.0: World Health Organization Disability Schedule 2.0.

[a] Refers to non-commercial use.

Affordability and Sustainability

Those seeking to select a measure for use in a research study or measurement-based care system face the challenge of identifying the most suitable tool from a wide array of choices. Psychometric systematic reviews and meta-analyses provide comparisons of measurement properties that can help with scale selection (eg., 30,31). Feasibility of use in a given target context is another important selection criterion, and within this, the scale's affordability (20,32). For example, clinical services or community-based providers may only have limited resources (if any) earmarked for measurement-based care (20). Within universities, research projects may require free measures due to limited funding (e.g., for student and trainee projects). For researchers and practitioners in lower- and middle-income countries (LMIC), the cost of commercially available measures may be prohibitive, especially if the tool is costed with reference to the purchasing power of users in high-income countries. On the grounds of equity and in order to build a diverse and inclusive evidence base, core metrics for research and measurement-based care should have the highest-possible level of affordability (28).

Many commonly used mental health measures and gold-standard diagnostic tools can be purchased for a fee, from commercial publishers. Within child and adolescent psychiatry for example, the widely used *Beck Youth Inventories Second Edition* (BYI-II; 22) measures are held by the educational publisher Pearson. As of March 2021, a starter kit of BYI-II tools including a manual and 25 paper-based inventory booklets could be purchased for $360.10 US dollars (33). Time, staff, and financial resources are usually required for the initial conception, design, and validation of a measurement scale (34), and developers may face additional costs associated with a scale's ongoing management and dissemination. By making a scale commercially available, developers shift a part of these costs onto users. User fees may also compensate for the provision of physical administration kits or further measure development although how exactly licence fees are utilized is rarely disclosed.

At the same time, a number of proprietary scales are provided at no cost to non-commercial users by individual developers, research groups, foundations, or organizations (Table 1), and several systematic reviews and repositories provide helpful catalogues of such freely available measures (20,32). However, "free" scales vary in the extent to which they are truly *free* or *open* for use, with varying restrictions placed on the user's right to modify, translate, or distribute them. Some tools are available in the public domain and may be used, modified, and distributed without limit and without seeking explicit permission to do so, while other tools may only be modified or translated with explicit permission from the developers (Table 1). While the exact terms of use are not always clearly articulated, the use of these scales typically does not require financial contributions from non-commercial users.

The free availability of some measurement scales might convey the impression that there is no marginal cost involved in controlling and coordinating their use. Yet, as highlighted above, ongoing support and monitoring is often needed to ensure that a measure is used appropriately, that is, in line with licencing terms; in ways that are methodologically sound; and under minimal risk of harm to respondents. Ongoing inputs may be needed to handle user queries, and to support and monitor other researchers' efforts to further validate or adapt the scale. While initial development is often funded through a research grant, activities related to the ongoing dissemination and oversight of measurement scales might not be explicitly funded.

Developers of freely available scales may currently be protected by the fragmentation of the field's attention across the many different existing options. However, if the common metrics movement were to succeed at centring measurement efforts around a small set of free common metrics, interest and support needs would likely grow exponentially. Where no continuous sources of funding are available, this may cause a dilemma for non-commercial providers struggling to meet this increased demand with the infrastructures and resources currently at their disposal. While the affordability of commercially licenced measures is comparatively poor from the users' point of view, these licencing models are sustainable for providers because they cover ongoing costs. This raises questions about how the field can ensure that a common metrics agenda centred around affordable measures is equitable and sustainable for users and providers alike.

Consistency and Adaptability

Common metrics initiatives such as COS aim to advance the field by integrating and harmonizing the evidence base, to enable comparisons, benchmarking, and syntheses of data across studies and clinical settings (5,9). Consistent use of a core set of scales (or items) across studies can facilitate the pooling of effect sizes in meta-analyses, and the pooling and linking of datasets as part of integrative data analysis projects that provide enhanced statistical power and allow for the investigation of new research hypotheses (35–37). In measurement-based care, the tracking of harmonized outcome indicators can enable comparisons between services and mental health systems, thus helping with the identification of best practice examples. In order to maintain comparability within a landscape of fixed measures, it is important that the recommended scales be used consistently and without substantial modification across studies and settings.

At the same time, many mental health scales have been developed in North America or Europe, in specific clinical settings, and the evidence base relating to their reliability and validity in other contexts (eg, LMIC; population-based surveys) is only gradually emerging. In turn, many measures that assess functional impairment or health-related quality of life were originally developed for use with populations in physical health care settings, or in non-clinical populations, and may require adaptation for meaningful use in mental health contexts.

More generally, there are important opportunities for strengthening measurement scales based on feedback from researchers, practitioners, and lived-experience experts. Currently, measurement is based on imperfect scales, some of which have been widely used for decades. For example, the Children's Depression Rating Scale−Revised (CDRS-R) (38) is the symptom measure that is most widely used across trials for youth depression treatments (39). However, a recent systematic review and evidence appraisal using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines shows that the evidence supporting the CDRS-R's measurement properties is weak and based on only six studies (40). In many cases, common metrics initiatives are therefore basing their recommendations on measures that are considered "good enough" rather than gold standard (2,5).

This is a compromise made to kick-start a common metrics movement and accelerate learning about optimal approaches to harmonization. It is based on the understanding that common metrics should be assessed *as a minimum*, but can be complemented flexibly with additional scales, which can help identify any issues, idiosyncrasies, or limitations of the recommended tools (41). To avoid a premature centring of measurement around tools that may later reveal important flaws, it is vital to pilot them in a variety of populations and contexts, and to adapt or exchange them if needed. For example, the implementation of ICHOM outcome sets is overseen by steering groups with a mandate to periodically review evidence from pilot studies and consider whether any measurement recommendations should be revised. At present, however, the fixed nature of most scales is a barrier to adaptation, as their revision or replacement may compromise the comparability of data collected across different time points or populations.

The proprietary nature of many existing scales is another barrier to adaptability. Commercially licenced measures often cannot be easily modified by users for applications in new populations or contexts. Control over the measure is often firmly centralized with the licence holder, which can mean foregoing opportunities for more dynamic and collective improvement efforts. While the scope for adaptation is comparatively low, this model provides maximum consistency, ensuring that only official versions of a scale are in circulation. Developers providing their tool at no cost often allow greater leeway for adaptation by considering collaborations, while at the same time striving to preserve the highest possible consistency. As discussed above, this two-fold effort may require considerable resources, especially if the relevant measurement scales find themselves at the heart of a common metrics movement. A third model, in which measures are fully in the public domain and open to modification and adaptation by anybody without control from a developer or provider meets criteria of affordability, sustainability, and adaptability, but provides no assurance of consistent use, which might undermine their utility as common metrics (Figure 1).

There is an obvious tension between the desire to enable iterative adaptation, on the one hand, and to preserve consistency in a scale's characteristics and administration, on the other hand. Similarly, there is a tension between desires for adaptation, and the licencing and copyright terms attached to proprietary measures, which may prohibit modifications. These tensions are difficult to reconcile within a framework of fixed and proprietary measures. Figure 1 visually illustrates our subjective understanding of the trade-offs involved in current models of measure development and dissemination in relation to affordability, sustainability, adaptability, and consistency. A "low" rating indicates that we perceive a comparatively small chance of providing the relevant attribute within the given development and dissemination model, or a high chance of providing only a limited amount. A "medium" rating indicates that we perceive there to be a medium chance of providing the attribute, or that it might be provided to some extent. A "high" rating indicates that we perceive a comparatively high chance of providing the relevant attribute within the given development and dissemination model. "Variable" ratings indicate that whether or not an attribute is likely to be achieved depends on the specificities of the given development and dissemination model.

Figure 1. Trade-Offs Between Measure Development and Dissemination Models in the Current Status Quo

	Paid-For Proprietary Measure	**Proprietary Measures Provided at no Cost**	**Measure Fully in the Public Domain**
Licencing Terms	Copyrighted	Copyrighted	Not Copyrighted
Affordability	Low	High	High
Sustainability	High	Low	High
Adaptability	Low	Variable	High
Consistency	High	Variable	Low

Where to Go from Here? A Three-Pronged Way Forward

It is our view that the inherent tensions between affordability and sustainability, consistency, and adaptability of fixed proprietary measures pose an important barrier to the successful implementation of the common metrics agenda in mental health. A focus on affordability without consideration of sustainability will likely be ineffective, as providers may find themselves unable to uphold free distribution models in the face of increasing demand for user support. In turn, a focus on sustainability without consideration of affordability would fall short on equity of access and the representativeness of the ensuing evidence base (28). As the common metrics agenda aims to generate a more inclusive, integrated, and higher-quality evidence base while still promoting harmonization, the competing needs for adaptation and consistency must also be considered. Hereafter, we outline a three-pronged approach as one possible avenue towards providing a better balance of affordability, sustainability, consistency, and adaptability in measure development and dissemination (Figure 2), and towards moving the development and dissemination of common metrics into the 21st century.

Figure 2. Trade-Offs Between Different Alternative Measure Development and Dissemination Models

	Funder-Led Coordination	Open Access Measure Hub	Item Bank
Affordability	High	High	Variable
Sustainability	Medium	Medium	Medium
Adaptability	Variable	Variable	High
Consistency	High	Medium	High

Changing the Funding and Dissemination Model for Proprietary Measures. Several funders, such as the National Institutes of Health (NIH) and the Wellcome Trust require that the funded research findings—and in some cases the underlying data—be shared via open access publishing or the deposition in accessible databases or repositories (28). Funders may want to extend such requirements to any measurement tools developed with their support, by either allowing researchers to budget for anticipated dissemination or maintenance costs in their initial grant applications, or by developing a mechanism for renewing funding for the maintenance and oversight of freely distributed metrics periodically. Funders might even establish systems whereby individual developers can transfer the responsibility for measure dissemination to the funder, who may then earmark resources and develop mechanisms for ongoing oversight. This model offers limited scope for dynamic scale adaptation, and is not inherently sustainable as it requires ongoing investment and commitment from funders. It does, however, provide relatively high affordability and scale consistency, and may enhance sustainability in the short term within a landscape of fixed proprietary scales.

Learning from Data Science: A Measure Hub Model. We propose that in the longer term, the field may need to shift its mindset from seeing measurement scales as fixed and proprietary, to thinking of measures as evolving tools or "code" that should be co-owned and co-created by the user community, including researchers, practitioners, and lived-experience experts. There may be opportunities for learning from pioneering work in the computer sciences

and statistics, and from open science approaches where it is thought that that openly available tools facilitate high-quality research practices and increase research efficiency (42).

For example, in the computer sciences, the Open Science Grid is a consortium of stakeholder communities (e.g., researchers, IT providers, software developers, educators) that share computing resources to advance scientific practice (43). Similarly, computer scientists and statisticians freely share code through platforms such as GitHub (https://github.com/). Developers can licence their code with a variety of licencing models that provide different levels of control and attribution, including permissive free software licences that impose minimal restrictions on use and distribution (e.g., BSD licences). The hub enables developers to monitor who is using their code and how it is being adapted. As such it provides a sophisticated and highly regulated model of version control, which may help reconcile adaptation and consistency—at least to a degree.

An open access measure hub in mental health could facilitate the sharing of non-commercial measurement scales, and their flexible adaptation for use in new settings and populations. The Wikiversity Evidence Based Assessment Portfolios already compile information related to mental health scales, which can be edited and expanded collaboratively by Wikiversity users (44). However, the portfolios do not currently host the scales themselves, or help with version control or the coordination of adaptation efforts. Other instrument repositories helpfully compile copies of free instruments where available, in addition to providing information on their measurement properties (20). However, existing repositories may not always be exhaustive or up-to-date, as highlighted by a recent review:

> "the measures contained in each repository often did not overlap and were not always updated with the latest versions of the measure. Repositories varied in how they selected measures to include; some required authors to self-submit and self-report on the measure's psychometric evidence, others gathered experts to recommend measures for inclusion. The dynamic nature of these repositories suggests that the landscape of freely available measures may shift quickly; however, in the absence of a single, coordinated effort to house pragmatic measures, these repositories are unlikely to keep pace with advancing science." (p.11) (20).

To "keep pace with advancing science," a more centralized and dynamic measure hub may be needed that can be updated by the user community itself. A centralized hub model can aid with sustainability by easing the cost and effort required for overseeing measure dissemination, where the hub provides a transparent track record of use, validation, and adaptation efforts. The onus for providing technical support and answering user queries could be shifted, at least partly, from individual developers onto the wider user community through open discussion boards and forums. Finally, a hub model might reduce the risk of inadequate or unauthorized use, by providing clear and transparent licencing terms. On the downside, while enabling more transparent version control, the hub model prioritizes adaptation over consistency, meaning that the number of different versions in circulation would likely increase, and comparability decline. There would also be no monetary compensation for a developer's initial development efforts and intellectual property, which may deter some from disseminating their tools in this way.

Moving from Fixed Measures to an Item Bank Model. While a centralized, open-access measure hub can make the adaptation and use of free measures easier and more dynamic, initial scale creation still lies with an individual developer or research team. Although a scale might be adaptable, it would remain a largely fixed tool. An alternative model that maximizes the potential for dynamic measure creation is a move from fixed scales to item banks and personalised assessment based on Item Response Theory (IRT) and Computerized Adaptive Testing (CAT) (45).

Item banks contain a large number of questionnaire items that, through use of IRT models, have been calibrated to assess a specific construct, such as depression or anxiety, at a certain level of difficulty (or severity). Using CAT, individual items can be selected flexibly to create tailored assessments that locate individuals more precisely and more rapidly on the construct continuum of interest than fixed scales that were created using classical test theory (CTT) (46,47). CAT-based item bank approaches enable consistent scoring across varying item combinations. As a result, it is possible to develop short forms that are tailored to specific populations, and still generate scores that are directly comparable (46), thus providing a way out of the adaptability-versus-consistency dilemma.

The National Institutes of Health's (NIH) Patient-Reported Outcomes Measurement Information System (PROMIS) is an item bank that was calibrated to capture patient-reported outcomes across a range of chronic health conditions. PROMIS items are assembled from existing measures via literature searches, reviewed through consultation with key stakeholders (including lived-experience experts), and subject to psychometric testing and validation. The PROMIS network is organized around several primary research sites, with a statistical coordinating centre managing the development and validation of items, and providing a data management and storage system.

To date, PROMIS has not replaced proprietary fixed measurement scales in much of child and youth anxiety and depression research, although relevant item banks and short forms are available (48). PROMIS measures were developed and calibrated as "non-disease-specific scales" to assess emotional health (including anxiety and depression) in individuals with chronic health conditions, rather than specifically for the purpose of clinical mental health assessment (p. 2196) (43). In addition, while printout short forms are available, personalised assessment via CAT requires digital administration, which is not (yet) feasible in all contexts. Although more sustainable than free measure distribution via individual developers, PROMIS does not have inbuilt sustainability as substantial inputs and resources by the PROMIS network, and hence continued external funding, are required to maintain the item bank. This is reflected in limitations to affordability, compared with freely distributed measures: While English and Spanish language fixed scales were available from PROMIS at no cost as of March 2021, translations did incur a distribution fee of $800 USD per instrument and language (https://www.healthmeasures.net). A review fee was charged for quality assuring new translations and for ensuring that they were harmonized across languages. The relatively slow uptake of the PROMIS measures for the purpose of clinical mental health assessment suggests a need to further examine the widespread feasibility of centring common metrics initiatives around item bank and IRT models.

In the meantime, IRT offers opportunities for harmonizing and comparing scores obtained from fixed legacy measures, which could help smoothen the transition from a fixed to a more dynamic measurement landscape. Using

IRT, items from existing fixed scales can be calibrated within the same measurement model used by an item bank, which allows linking the scores of legacy scales to a common underlying scoring metric. The scores of different scales can then be compared via crosswalk tables (see prosetta-stone.org) or online resources (see common-metrics.org). Relevant studies have been conducted using PROMIS as a common metric for several common depression and anxiety scales as part of the PROsetta Stone project (50–53). Once a linkage is established, comparability can be preserved when there is a need to move from one fixed scale to another. The continued development of such models and the linkage of additional scales may eventually offer a perspective for creating common metrics that are independent from individual instruments, without requiring the use of a single common instrument, and without sacrificing the comparability of legacy data. Funders could support this transition by requiring that any newly developed or adapted fixed scales be calibrated onto an existing measurement model.

Looking Ahead

To overcome the current state of data fragmentation in mental health research and to promote the creation of a consistent, diverse, and inclusive evidence base, it is important to centre research efforts around common metrics that are affordable to avoid the emergence of a new line of fragmentation driven by cost. The latter would also undermine efforts to enhance the quality of the evidence base, by making the most suitable tools the purview of a comparatively small group of researchers and practitioners with the necessary monetary resources.

Recent common metrics and Core Outcome Set initiatives in anxiety and depression research (5,13,15,16) have recommended measures like the Generalized Anxiety Disorder 7-item Scale (GAD-7; 24), the Patient Health Questionnaire (PHQ-9; 25); the World Health Organization Disability Schedule 2.0 (WHODAS 2.0) 12-item short form (26), and the Revised Children's Anxiety and Depression Scale 25-item short version (54). While these were selected for being freely available at the time of selection (amongst other criteria), there is no guarantee that they will be provided at no cost indefinitely, especially if developers or providers begin to feel overburdened by increasing demand for support. While it was considered that these measures were "good enough" (2) for now, they may reveal important limitations, once piloted across a greater variety of contexts and settings. To base the common metrics movement on the best-possible measurement scales, it is vital that adaptation, modification, and tailoring to specific populations is possible, without undermining the overarching aim of harmonization. The future of the common metrics agenda may lie in moving beyond fixed proprietary measurement scales towards models that provide greater scope for dynamic adaptation and tailoring, while maintaining the necessary degree of consistency. Item banks, IRT, and CAT will likely play a central role in this effort.

As a field, we need to rethink how measure development and dissemination can be organized within a common metrics framework. No single model is likely to resolve the outlined tensions between affordability, sustainability, adaptability, and consistency today. We have discussed a three-pronged way forward that involves revisiting current models for the dissemination of fixed proprietary measures, and considering opportunities for the development of more dynamic assessments. Rather than thinking of measures as 20th-century manuscripts or instruments, it may be

helpful to start thinking of them as the equivalent of 21st-century tools or computation code that can be co-owned and co-created by the wider research community with support from dedicated infrastructure, coordinating bodies, or funders. Funders have already started to agree the key building blocks of a common metrics toolbox (55). The next step is to help support the wider mental health science community to start building from these to create free and flexible metrics in ways that are sustainable and also preserve comparability and consistency.

References

1. Harding KJK, Rush AJ, Arbuckle M, Trivedi MH, Pincus HA. Measurement-based care in psychiatric practice: A policy framework for implementation. J Clin Psychiatry. 2011 Aug 15;72(8):1136–43.
2. The Lancet Psychiatry. A good enough measure. The Lancet Psychiatry. 2020 Oct;7(10):825.
3. Santor DA, Gregus M, Welch A. FOCUS ARTICLE: Eight Decades of Measurement in Depression. Meas Interdiscip Res Perspect. 2006 Jul;4(3):135–55.
4. Mew EJ, Monsour A, Saeed L, Santos L, Patel S, Courtney DB, et al. Systematic scoping review identifies heterogeneity in outcomes measured in adolescent depression clinical trials. J Clin Epidemiol. 2020 Oct;126:71–9.
5. Krause KR, Chung S, Adewuya AO, Albano AM, Babins-Wagner R, Birkinshaw L, et al. International consensus on a standard set of outcome measures for child and youth anxiety, depression, obsessive-compulsive disorder, and post-traumatic stress disorder. The Lancet Psychiatry. 2021;8(1):76–86.
6. Fried EI. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. J Affect Disord. 2017 Jan;208(July 2016):191–7.
7. Newson JJ, Hunter D, Thiagarajan TC. The Heterogeneity of Mental Health Assessment. Front Psychiatry. 2020 Feb 27;11:1–24.
8. Bear HA, Edbrooke-Childs J, Norton S, Krause KR, Wolpert M. Systematic Review and Meta-analysis: Outcomes of Routine Specialist Mental Health Care for Young People With Depression and/or Anxiety. J Am Acad Child Adolesc Psychiatry. 2020;59(7):810–41.
9. Szatmari P, Offringa M, Butcher NJ, Monga S. Counting What Counts: The Case for Harmonized Outcomes in Child and Youth Mental Health Research. J Am Acad Child Adolesc Psychiatry. 2019;58(7):656–8.
10. Monga S, Offringa M, Butcher NJ, Szatmari P. From Research to Practice: The Importance of Appropriate Outcome Selection, Measurement, and Reporting in Pediatric Mental Health Research. J Am Acad Child Adolesc Psychiatry. 2020;59(4):497–500.
11. Clarke M, Williamson PR. Core outcome sets and systematic reviews. Syst Rev. 2016;5(1):11.
12. Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, et al. The COMET Handbook: Version 1.0. Trials. 2017;18(Suppl 3):1–50.
13. Obbarius A, van Maasakkers L, Baer L, Clark DM, Crocker AG, de Beurs E, et al. Standardization of health outcomes assessment for depression and anxiety: recommendations from the ICHOM Depression and Anxiety Working Group. Qual Life Res. 2017;26(12):3211–25.
14. Chevance A, Ravaud P, Tomlinson A, Le Berre C, Teufer B, Touboul S, et al. Identifying outcomes for

depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. The Lancet Psychiatry. 2020 Aug;7(8):692–702.

15. Wolpert M. Funders agree first common metrics for mental health science [Internet]. Linkedin. 2020 [cited 2021 Mar 5]. Available from: https://www.linkedin.com/pulse/funders-agree-first-common-metrics-mental-health-science-wolpert/

16. UNICEF. Measurement of Mental Health Among Adolescents at the Population Level (MMAP): An Overview [Internet]. 2019 [cited 2021 Mar 5]. Available from: https://data.unicef.org/wp-content/uploads/2018/02/Mental-health-measurement_MMAP_overview_UNICEF.pdf

17. Boswell JF, Kraus DR, Miller SD, Lambert MJ. Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. Psychother Res. 2015 Jan 2;25(1):6–19.

18. Kotte A, Hill KA, Mah AC, Korathu-Larson PA, Au JR, Izmirian S, et al. Facilitators and Barriers of Implementing a Measurement Feedback System in Public Youth Mental Health. Adm Policy Ment Heal Ment Heal Serv Res. 2016 Nov 21;43(6):861–78.

19. Whiteside SPH, Sattler AF, Hathaway J, Douglas KV. Use of evidence-based assessment for childhood anxiety disorders in community practice. J Anxiety Disord. 2016;39:65–70.

20. Becker-Haimes EM, Tabachnick AR, Last BS, Stewart RE, Hasan-Granier A, Beidas RS. Evidence Base Update for Brief, Free, and Accessible Youth Mental Health Measures. J Clin Child Adolesc Psychol. 2020 Jan 2;49(1):1–17.

21. Leval PN. Campbell as Fair Use Blueprint? Washingt Law Rev. 2015;90(2):597.

22. Beck JS, Beck AT, Jolly J, Steer RA. Beck Youth Inventories™ - Second Edition For Children and Adolescents (BYI-II). 2nd ed. San Antonio, TX: Pearson; 2005.

23. Garralda ME, Yates P, Higginson I. Child and adolescent mental health service use. HoNOSCA as an outcome measure. Br J Psychiatry. 2000;177:52–8.

24. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A Brief Measure for Assessing Generalized Anxiety Disorder. Arch Intern Med. 2006 May 22;166(10):1092.

25. Spitzer RL, Kroenke K, Williams JBW. Validation and utility of a self-report version of PRIME-MD: The PHQ Primary Care Study. J Am Med Assoc. 1999;282(18):1737–44.

26. Üstün T, Chatterji S, Kostanjsek N, Rehm J, Kennedy C, Epping-Jordan J, et al. Developing the World Health Organization disability assessment schedule 2.0. Bull World Health Organ. 2010;88(11):815–23.

27. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item

banks: 2005-2008. J Clin Epidemiol. 2010;63(11):1179–94.

28. Hays RD, Weech-Maldonado R, Teresi JA, Wallace SP, Stewart AL. Commentary: Copyright Restrictions Versus Open Access to Survey Instruments. Med Care. 2018;56(2):107–10.

29. Retraction Watch. Psychology paper retracted after creators of tool allege "serious breach of copyright" [Internet]. 2020 [cited 2021 Jan 5]. Available from: https://retractionwatch.com/2020/12/29/psychology-paper-retracted-after-creators-of-tool-allege-serious-breach-of-copyright/

30. Youngstrom EA, Egerton GA, Genzlinger J, Freeman LK, Rizvi SH, Van Meter A. Improving the global identification of bipolar spectrum disorders: Meta-analysis of the diagnostic accuracy of checklists. Psychol Bull. 2018 Mar;144(3):315–42.

31. Stockings E, Degenhardt L, Lee YY, Mihalopoulos C, Liu A, Hobbs M, et al. Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. J Affect Disord. 2015 Mar;174:447–63.

32. Beidas RS, Stewart RE, Walsh L, Lucas S, Downey MM, Jackson K, et al. Free, brief, and validated: Standardized instruments for low-resource mental health settings. Cogn Behav Pract. 2015;22(1):5–19.

33. Pearson Assessments. Beck Youth Inventories Second Edition [Internet]. [cited 2021 Mar 2]. Available from: https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Personality-%26-Biopsychosocial/Beck-Youth-Inventories-%7C-Second-Edition/p/100000153.html#

34. Mokkink LB, Prinsen CA, Patrick D, Alonso J, Bouter LM, de Vet HC, et al. COSMIN Study Design checklist for Patient-reported outcome measurement instruments [Internet]. 2019. Available from: https://gut.bmj.com/content/gutjnl/70/1/139/DC1/embed/inline-supplementary-material-1.pdf?download=true

35. Hussong AM, Curran PJ, Bauer DJ. Integrative data analysis in clinical psychology research. Annu Rev Clin Psychol. 2013;9:61–89.

36. Hofer SM, Piccinin AM. Integrative Data Analysis Through Coordination of Measurement and Analysis Protocol Across Independent Longitudinal Studies. Psychol Methods. 2009;14(2):150–64.

37. Bainter SA, Curran PJ. Advantages of Integrative Data Analysis for Developmental Research. J Cogn Dev. 2015;16(1):1–10.

38. Poznanski E, Mokros H. Children's Depression Rating Scale–Revised (CDRS-R). Los Angeles, CA: Western Psychological Services; 1996.

39. Krause KR, Bear HA, Edbrooke-Childs J, Wolpert M. Review: What Outcomes Count? A Review of Outcomes Measured for Adolescent Depression Between 2007 and 2017. J Am Acad Child Adolesc Psychiatry. 2019;58(1):61–71.

40. Stallwood E, Monsour A, Rodrigues C, Monga S, Terwee C, Offringa M, et al. Systematic Review: The Measurement Properties of the Children's Depression Rating Scale−Revised in Adolescents With Major Depressive Disorder. J Am Acad Child Adolesc Psychiatry. 2021;60(1):119–33.

41. Patalay P, Fried EI. Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. J Child Psychol Psychiatry. 2020 Sep 28;7:jcpp.13333.

42. Bartling S, Friesike S. Opening Science. Opening Science. Cham, CH: Springer International Publishing; 2014.

43. Altunay M, Avery P, Blackburn K, Bockelman B, Ernst M, Fraser D, et al. A Science Driven Production Cyberinfrastructure—the Open Science Grid. J Grid Comput. 2011;9(2):201–18.

44. Wikiversity. Evidence-based assessment/Assessment portfolios V2 [Internet]. 2021 [cited 2021 Feb 24]. Available from: https://en.wikiversity.org/wiki/Evidence-based_assessment/Assessment_portfolios_V2#Instructions_for_Editing_a_Portfolio

45. Thissen D, Steinberg L. Item response theory. Sage Handb Quant methods Psychol. 2009;148–77.

46. Amtmann D, Cook KF, Jensen MP, Chen W-H, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010 Jul;150(1):173–82.

47. Revicki DA, Cella DF. Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. Qual Life Res. 1997;6(6):595–600.

48. Irwin DE, Stucky B, Langer MM, Thissen D, DeWitt EM, Lai J-S, et al. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. Qual Life Res. 2010;19(4):595–607.

49. DeWalt DA, Gross HE, Gipson DS, Selewski DT, DeWitt EM, Dampier CD, et al. PROMIS® pediatric self-report scales distinguish subgroups of children within and across six common pediatric chronic health conditions. Qual Life Res. 2015 Sep 26;24(9):2195–208.

50. 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. Psychol Assess. 2014;26(2):513–27.

51. Kaat AJ, Newcomb ME, Ryan DT, Mustanski B. Expanding a common metric for depression reporting: linking two scales to PROMIS® depression. Qual Life Res. 2017;26(5):1119–28.

52. Kaat AJ, Kallen MA, Nowinski CJ, Sterling SA, Westbrook SR, Peters JT. PROMIS® Pediatric Depressive Symptoms as a Harmonized Score Metric. J Pediatr Psychol. 2020;45(3):271–80.

53. Schalet BD, Cook KF, Choi SW, Cella D. Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. J Anxiety Disord. 2014;28(1):88–96.

54. Ebesutani C, Reise SP, Chorpita BF, Ale C, Regan J, Young J, et al. The Revised Child Anxiety and

Depression Scale-Short Version: Scale reduction via exploratory bifactor modeling of the broad anxiety factor. Psychol Assess. 2012;24(4):833–45.

55. International Alliance of Mental Health Research Funders. Year in Review: Highlights from the IAMHRF's work in the year 2019-2020 [Internet]. 2020. Available from: https://iamhrf.org/sites/iamhrf.org/files/uploads/page/files/iamhrf_highlights_2020_v1.pdf