Implementation measurement in global mental health: Results from a modified Delphi panel and investigator survey

Christopher G. Kemp[1] [ORCID], Kristen Danforth[2], Luke Aldridge[3], Laura K. Murray[3] and Emily E. Haroz[1,3] [ORCID]

[1]Department of International Health, Johns Hopkins University, Baltimore, MD, USA; [2]University of Washington, Seattle, WA, USA and [3]Department of Mental Health, Johns Hopkins University, Baltimore, MD, USA

Abstract

Limited guidance exists to support investigators in the choice, adaptation, validation and use of implementation measures for global mental health implementation research. Our objectives were to develop consensus on best practices for implementation measurement and identify strengths and opportunities in current practice. We convened seven expert panelists. Participants rated approaches to measure adaptation and validation according to appropriateness and feasibility. Follow-up interviews were conducted and a group discussion was held. We then surveyed investigators who have used quantitative implementation measures in global mental health implementation research. Participants described their use of implementation measures, including approaches to adaptation and validation, alongside challenges and opportunities. Panelists agreed that investigators could rely on evidence of a measure's validity, reliability and dimensionality from similar contexts. Panelists did not reach consensus on whether to establish the pragmatic qualities of measures in novel settings. Survey respondents ($n = 28$) most commonly reported using the Consolidated Framework for Implementation Research Inner Setting Measures ($n = 9$) and the Program Assessment Sustainability Tool ($n = 5$). All reported adapting measures to their settings; only two reported validating their measures. These results will support guidance for implementation measurement in support of mental health services in diverse global settings.

Impact statement

Growth in the need for rigorous implementation science in global mental health research has outpaced the development and validation of pragmatic tools to measure implementation processes and outcomes in diverse global settings. Of the few implementation measures that are currently in use, essentially all were developed for use in high-income settings, and few have been psychometrically assessed or validated. Our objectives were to (1) bring together a panel of experts and build consensus around best practices for implementation measurement in diverse global settings and (2) survey investigators applying these measures to identify strengths and opportunities in current practice. The results will support guidance for use by investigators planning to quantitatively measure implementation process and outcomes in diverse global settings. This guidance could facilitate novel, rigorous and replicable implementation research in areas of high need.

Cambridge Prisms

CAMBRIDGE UNIVERSITY PRESS

Introduction

Mental, neurological and substance-use (MNS) disorders are the leading causes of disability globally, yet most people in need of treatment for MNS disorders never receive care (Thornicroft et al., 2017; Pathare et al., 2018; Vos et al., 2020). Effective, affordable, scalable and sustainable services are needed to bridge this global gap (Lancet Global Mental Health Group et al., 2007). A broad range of preventive and treatment interventions for high-burden MNS conditions have demonstrated promising cost-effectiveness in both high- and low-resource settings (Patel et al., 2016); in response, researchers and funders alike have called for an increased scientific focus on strengthening intervention implementation and scale-up, particularly in low- and middle-income countries (LMICs), through the application of the methods of implementation science (Betancourt and Chambers, 2016). The primary aim of implementation science is to design and test ways to promote and sustain the delivery of evidence-based practices in routine healthcare (Eccles and Mittman, 2006). These *implementation strategies* target specific aspects of the environment of service delivery, or of the intervention providers or of the intervention itself, all with the goal of improving uptake and sustainment. Implementation success is assessed through a

range of implementation outcomes, including acceptability, adoption, appropriateness, cost, feasibility, fidelity, penetration and sustainability (Proctor et al., 2011). For example, if unhelpful attitudes or beliefs among clinic staff are thought to be hindering implementation of evidence-based mental health care, the use of peer influencers or opinion leaders might be considered as an implementation strategy to improve provider acceptance of mental health services. Application of implementation science methods to the field of global mental health has grown rapidly in recent years (Wagenaar et al., 2020).

This growth has outpaced the development and validation of pragmatic tools for implementation measurement in diverse global settings. As with any science, valid measurement is critical to the utility and reproducibility of implementation research (Lewis et al., 2015). For example, many implementation studies begin with an assessment of the multi-level contextual determinants of implementation effectiveness (Damschroder et al., 2009). These determinants can inform the choice of implementation strategies; they are also useful for understanding the process of implementation and they may moderate or mediate intervention effects (Waltz et al., 2019). Measurement of implementation outcomes is also critical to judging the effectiveness of implementation strategies. While some implementation constructs may be *manifest*, or measured through observable indicators (e.g., rate of provider serviced delivery as an indicator of penetration) (Willmeroth et al., 2019), many are *latent*, implying some level of self-report (e.g., provider acceptability). Many quantitative measures of latent implementation constructs exist and have been identified and catalogued through systematic review; relatively few, however, have been assessed for validity or have documented strong psychometric properties, though the number of measures with strong psychometric properties is increasing (Khadjesari et al., 2020; Mettert et al., 2020). Even fewer measures have been assessed for their pragmatic qualities, including burden, length, reliability and sensitivity to change (Hull et al., 2022). Importantly, almost all extant, validated, pragmatic, quantitative implementation measures were developed for use in high-income countries (Lewis et al., 2015). These implementation measures – and their corresponding theories, models and frameworks – may need to be appropriately translated, adapted and validated for use in diverse global contexts (Means et al., 2020).

To date, most implementation studies by global mental health researchers have relied exclusively on qualitative assessment, with relatively few using quantitative implementation measures (Wagenaar et al., 2020). Though qualitative methods are a crucial part of implementation science, valid quantitative measurement allows for larger studies and improves study rigor and reproducibility (Palinkas et al., 2011; Palinkas, 2014). Investigators have several factors to consider when choosing quantitative measures for use – in addition to whether an appropriate measure exists – including different aspects of measure validity and reliability, as well as each measure's pragmatic qualities (e.g., length, cost) (Powell et al., 2017). Given that almost all existing implementation measures were developed for use in high-resource settings, global mental health researchers must carefully consider the validity and appropriateness of each measure in their setting. There are several distinct approaches available for establishing validity and other measure characteristics in novel settings (Boateng et al., 2018). Table 1 describes these characteristics and approaches in detail and notes which approaches are designed to assess which characteristics. For example, cross-cultural validity can be established using translation, back-translation, expert advice and pre-testing.

Limited guidance exists to support global mental health services investigators in the choice and use of quantitative implementation

measures – or the choice and use of approaches to adapt and validate those measures. Our objectives in this project were to (1) bring together a panel of experts to better understand and develop consensus on best practices for implementation measurement, with a particular focus on mental health implementation research in LMICs, and (2) survey investigators applying these measures to identify strengths and opportunities in current practice.

Methods

Expert panel

Participants

We used purposive sampling to select and invite a panel of experts at the intersection of implementation science, psychometrics and global mental health, starting from a list generated by members of the study team. Specifically, we approached experts in our extended professional networks who we knew had experience with developing, adapting or validating implementation measures for use in global mental health research. We recruited eight panel members (see Supplementary Material for a full list of panel participants). One panel member withdrew between the first and second panel discussions.

Delphi process

The goal of our modified Delphi process was to develop consensus among the panel members on: (1) prioritization of different types of measure validity, reliability and pragmatic qualities for assessment and confirmation when using measures under different circumstances and in different settings (see Table 1 for definitions of each quality); (2) feasibility and utility of different measure validation approaches (see Table 1 for definitions of each approach) and (3) a minimal set of validation approaches for use when applying implementation measures in new contexts and settings. We followed the steps of a conventional Delphi process, including an exploratory phase, a first round of quantitative questionnaires, analysis/summation and results discussion (Avella, 2016). A preliminary discussion was held in March 2020 to orient panelists to the Delphi process. Questionnaires were then distributed and completed electronically. Questionnaire responses were aggregated and anonymized, and summary statistics of responses were presented to the panel. Following the distribution of the questionnaire analysis, available panel members were convened virtually to review the results and, if possible, achieve consensus on recommendations.

Questionnaires included three sections (see Supplementary Material). In the first section, panel members were given different measurement scenarios (e.g., use of an implementation measure developed in a US context to assess the same construct in a novel, lower-resource context) and were asked which types of measurement characteristics (e.g., different types of validity, reliability or pragmatic qualities; Table 1) need to be established prior to measure use in a novel context. In the second section, panel members rated distinct validation strategies (e.g., informal expert elicitation, pilot survey with subsequent real-world outcomes; Table 1) on nine dimensions of rigor, feasibility and resource intensiveness. Finally, in the third section panel members proposed a minimal set of validation strategies that researchers could use under most circumstances when applying an implementation measure in a diverse new setting.

One author (KD) had access to the questionnaire responses and interview data and completed all analyses (Linstone and Turoff, 1975). To maintain confidentiality and promote the rigor of the

Table 1. Implementation measure characteristics mapped to measure assessment approaches

Study Designs	Summary	Validity — Content validity: Substantive content validity (Is the measure reflective of the construct of interest?)	Validity — Content validity: Discriminant content validity (Is the measure non-reflective of other constructs?)	Validity — Criterion validity: Predictive validity (Does the measure predict future outcomes?)	Validity — Criterion validity: Concurrent validity (Does the measure correlate with concurrent gold-standard criterion measurements?)	Validity — Construct validity: Convergent validity (Does the measure correlate with other measures of the same construct?)	Validity — Construct validity: Discriminant validity (Does the measure correlate with other measures of different constructs?)	Validity — Construct validity: Know-groups differentiation (How does the measure behave across two or more known groups?)	Validity — Construct validity: Correlation analysis (How does the measure correlate with existing measures?)	Validity — Construct validity: Cross-cultural validity (Ensuring a tool is adapted for a new context and performs adequately)	Dimensionality (Does the measure capture the hypothesized structure of the construct?)	Reliability: Internal consistency (Do the scale items co-vary relative to their sum score?)	Reliability: Test-retest reliability (Are participant performances repeatable?)	Pragmatic Qualities — Stakeholder facing — Acceptable: Acceptable	Pragmatic Qualities — Stakeholder facing — Acceptable: Offers relative advantage over existing methods	Pragmatic Qualities — Stakeholder facing — Easy: Completed with ease	Pragmatic Qualities — Stakeholder facing — Compatible: Appropriate	Pragmatic Qualities — Stakeholder facing — Useful: Fits organizational activities	Pragmatic Qualities — Stakeholder facing — Useful: Informs clinical or organizational decision-making	Pragmatic Qualities — Objective — Acceptable: Cost	Pragmatic Qualities — Objective — Easy: Uses accessible language	Pragmatic Qualities — Objective — Easy: Assessor burden (training)	Pragmatic Qualities — Objective — Easy: Assessor burden (interpretation)	Pragmatic Qualities — Objective — Easy: Length
Informal expert elicitation	Ask a few researchers to consider whether one or more measure's items seem representative of each construct's theoretical content.	X	X (if including multiple measures)																					
Formal expert or target population elicitation (e.g. Delphi)	Ask a large group of researchers and practitioners to rate the extent to which items reflect the constructs they were intended to measure. Perform formal quantitative analysis.	X	X (if including multiple measures)																					
Translation, back-translation, expert advice, pre-testing with cross-sectional or repeat-measures survey	Use professional services and expert advice to translate and back-translate measure to/from target language and cultural context, pre-test with target population									X	X	X	X (with repeated measurement)											
Survey with sample of target population, using novel measures and established measures of similar or dissimilar constructs	Conduct a cross-sectional or repeat-measures survey with target population, including new measure(s) as well as established measures of similar and/or dissimilar constructs.				X (with gold standard measure)	X (with measure of similar construct)	(with measure of dissimilar construct)	X (pre-stratified target population)	X		X	X	X (with repeated measurement)											
Survey with sample of target population, using novel measures and subsequent real-world outcomes (e.g., EBP adoption, patient health)	Conduct a cross-sectional or repeat-measures survey with target population, including new measure(s) and referent EBP at t=0, and observe relevant outcomes at t=1			X	X (with gold standard measure)	X (with measure of similar construct)	(with measure of dissimilar construct)	X (pre-stratified target population)	X (if including other measures)		X	X	X (with repeated measurement)											
Vignette-based approach, using novel measures and established measures of similar or dissimilar constructs	Use measures with target population and randomly assigned vignette that systematically varies with respect to construct of interest.				X	X (with measure of similar construct)	(with measure of dissimilar construct)	X (pre-stratified target population)	X		X	X	X (with repeated measurement)											
Process evaluation of measure use	e.g., time-motion costing, qualitative study of stakeholder acceptability of measurement use													X	X	X	X	X	X	X	X	X	X	X

process, no identifying information was shared with other members of the research team or expert panel. Results draw from all questionnaire and interview responses as well as discussion during the second-round call. CK moderated and LA attended, but did not contribute to, both rounds of panel discussion.

The aim was to achieve a reasonable degree of consensus among panel members. No *a priori* target for degree of consensus was set for this study, and a full consensus-based approach was not pursued. This was done for reasons of appropriateness and feasibility; in particular, there are only a small number of experts at the intersection of global mental health and implementation measurement worldwide, and ongoing travel restrictions and social distancing measures related to the COVID-19 pandemic meant in-person consensus-building activities were impossible at the time. Though we did not use a quantitative threshold (e.g., calculating an agreement statistic or a formal vote) to assess consensus, we did bring the expert panel together for a Zoom-based discussion of the summary of their questionnaire results, with a particular focus on areas of divergence. Panel members agreed with the synthesis of results and concluded that the rankings of results within each subsection were acceptable and reflected their judgement.

Investigator survey

Participants

We also conducted a survey of global mental health researchers to understand current practice in implementation measurement. We searched NIH RePORTER and the Grand Challenges Canada website on May 18, 2020, for descriptions of funded implementation research studies related to mental health services in LMIC settings (see Supplementary Material for the NIH RePORTER search strategy). The names and contact information for the lead principal investigator for each study, as well as study descriptions, were abstracted into a sampling frame. One of three authors (C.G.K., K.D., L.A.) screened each study and associated principal investigator for inclusion; studies were excluded if they were not conducted in an LMIC or were not related to mental health. We contacted all remaining principal investigators and invited them to participate in a structured online survey related to the measurement of implementation processes and outcomes in their study. Principal investigators could also nominate a study team member or collaborator – someone who was directly involved in the implementation measurement component of the study – to participate in their place. Between NIH RePORTER, Grand Challenges Canada and this snowball sampling approach, we anticipated reaching most investigators with experience leading formal global mental health implementation research. Contacted investigators were sent a reminder email if they did not initially respond to the online questionnaire within a 2-week period, and a final reminder was sent 2 weeks later. Survey recruitment and data collection occurred from July to November 2020.

Survey measures

We designed the survey to assess: (1) the scope and nature of global mental health implementation research conducted by each investigator, (2) the range of implementation process and outcome measures used by investigators across any of their implementation studies and (3) the study setting, population, sample size, types of measure adaptation or validation used if any, assessment of measure performance and any recommendations for measure improvement.

Analysis

Categorical responses were summarized using simple descriptive statistics at the level of the respondent. Open-text responses were reviewed for recurring themes or approaches to adaptation and validation.

Research ethics

The Human Subjects Division of the University of Washington determined that both components of this study qualified for exemption status under 45 CFR 46.101 (b).

Results

Expert panel

Section 1: Measure characteristics

There was substantial concordance across panel members indicating it was reasonable to rely on evidence of most measure characteristics that had been established in similar contexts (e.g., another low-resource setting) without needing to establish those characteristics in every new setting (Supplementary Material, Section 1). This was true for all types of measure validity, reliability and dimensionality, except for cross-cultural validity (i.e., adequate adaptation for and performance in a new context), which was judged important to be established in each new setting. In contrast, there was limited agreement on the need to establish the pragmatic qualities of measures in each new setting. Though qualities like measure cost, length, ease of completion and assessor burden were judged to be unnecessary to be established in new settings if already established in similar settings, qualities related to how the measure would be used (e.g., whether it would inform decision-making, whether it fit with organizational activities) were felt to be important to establish in each new setting.

Panel members were then asked whether it was ever possible to rely on evidence of measure characteristics that had been established in other settings, even settings that were substantially different (e.g., high-income country). Respondents indicated that if investigators established the face validity of an implementation measure in a new setting – for example, through informal expert review and a small pilot use with confirmatory factor analysis – it would not then be necessary to conduct an intensive validation process. Respondents suggested that because implementation measures were not used directly to guide patient care, the stakes were lower than for other measures (e.g., diagnostic or screening tools), and correspondingly the bar for validation was lower.

Panel members were also asked about how they would choose between different hypothetical implementation measures based on their pragmatic qualities, assuming the hypothetical measures were equally valid. Respondents scored nearly all pragmatic qualities as important in making this decision, though acceptability, ease of completion, cost and language accessibility were rated as the most important qualities that would be considered (Table 2). In follow-up conversations with panel members, nearly all highlighted measure length as a key issue with current implementation measures, raising concerns related to respondent fatigue, assessor fatigue and artificial inflation of internal consistency. Respondents also felt that the results from most currently available measures were difficult to interpret, and that this was holding back their use and applicability. They suggested that the inclusion of quantitative thresholds and other guidance on how to judge what measure scores "mean" would be beneficial.

Table 2. Delphi panel pragmatic qualities importance ratings

	Average score
Acceptable (*Do users like the measure?*)	3.8
Completed with ease (*How hard is the measure to complete?*)	3.8
Cost (*Is the measure free to use?*)	3.8
Uses accessible language (*What is the reading level of the measure?*)	3.8
Appropriate (*Does use of the measure interfere with service implementation?*)	3.4
Length (*How many items does the measure have?*)	3.2
Informs clinical or organizational decision-making (*Are the measure findings actionable?*)	3
Fits organizational activities (*Does the measure map to actual services?*)	2.8
Assessor burden (training) (*How much training is required to learn how to administer the measure?*)	2.8
Assessor burden (interpretation) (*Does the measure have clear cut-offs, instructions for handling missing data and generating summary scores?*)	2.8
Offers relative advantage over existing methods (*Is the measure better than other approaches to assessment of the same construct?*)	2.4

Table 3. Investigator survey respondent characteristics (*n* = 28)

	Total (%)
Country of residence (%)	
Ethiopia	1 (3.6)
India	2 (7.1)
Lebanon	1 (3.6)
Nepal	1 (3.6)
Netherlands	1 (3.6)
Nigeria	1 (3.6)
Pakistan	1 (3.6)
South Africa	1 (3.6)
Ukraine	1 (3.6)
United Kingdom	1 (3.6)
United States	17 (60.7)
Organization type (%)[a]	
University/Academic	23 (82.1)
Non-profit/NGO	7 (25.0)
Government	1 (3.6)
Healthcare setting	1 (3.6)
Role (%)[a]	
Academic/research	27 (96.4)
Clinical service delivery	4 (14.3)
Program implementation	5 (17.9)
Policymaking	1 (3.6)
Time working in mental health	
0–4 years	3 (10.7)
5–14 years	12 (42.9)
15+ years	13 (46.4)
Mental health implementation studies conducted (mean (SD))	2.21 (1.37)

[a]≥1 response per participant possible.

Section 2: Validation strategies

Respondents identified a trade-off between the rigor of different validation approaches and their resource-intensiveness (Supplementary Material, Section 2). The two survey-based validation strategies, one using other established measures and the other using subsequent real-world outcomes for validation, were judged to be the most rigorous as well as the most expensive and time-consuming. Respondents rated the two forms of expert elicitation (informal and formal) as moderately or highly feasible and inexpensive, but there was no agreement on the assumed rigor of the results. Translation/back-translation scored consistently and moderately on all dimensions. Respondents disagreed most about the vignette-based strategy; they did not agree on the amount of time and resources required, nor whether it was feasible to develop vignettes that could provide high-confidence results in diverse low-resource settings. One respondent cautioned that developing good vignettes for community mental health programs could be hampered by the fact that these services are often uncommon in low-resource settings, and thus there is no "gold standard" program to which one can refer. Instead, vignettes must use hypothetical examples that take longer to explain and may produce unreliable results.

Section 3: Package of validation strategies

Translation/back-translation was the most frequently recommended strategy followed by informal expert elicitation. No other strategy was recommended by more than two respondents. Several respondents struggled with the tension between cost and rigor and wondered whether a minimal set of validation strategies might be feasible in most situations but ultimately insufficient for establishing validity. Most respondents suggested using a combination of validation strategies was the most appropriate approach; nearly all respondents argued that strategies should be "fit for purpose" and only as rigorous and complex as necessary. Respondents also debated the most appropriate approach to disseminate guidance on implementation measurement to mental health services researchers across diverse global settings. One respondent argued for the provision of step-by-step guidance, while another cautioned against offering overly prescriptive guidance to LMIC-based investigators.

Complete Delphi panel results are presented in the Supplementary Material.

Table 4. Implementation measure usage and adaptation/validation approaches

	CFIR (n = 7)	PSAT (n = 5)	AIM, IAM, FIM (n = 5)	AMHR/mhIST (n = 4)	EBPAS (n = 4)	ORIC (n = 3)
Implementation phase of application	Pre (n = 5); mid (n = 4); post (n = 2)	Pre (n = 2); mid (n = 4); post (n = 1)	Pre (n = 2); mid (n = 5); post (n = 1)	Pre (n = 1); mid (n = 2); post (n = 2)	Pre (n = 4); mid (n = 2); post (n = 1)	Pre (n = 2); mid (n = 1)
Measure use	Contextual determinant/process (n = 6); outcome (n = 1)	Contextual determinant/process (n = 3); outcome (n = 3)	Contextual determinant/process (n = 3); outcome (n = 2)	Contextual determinant/process (n = 1); outcome (n = 3)	Contextual determinant/process (n = 4)	Contextual determinant/process (n = 3)
Level of application	Client (n = 4); provider (n = 6); organizational (n = 3); policy (n = 2)	Provider (n = 5); organizational (n = 3)	Client (n = 1); provider (n = 4); organizational (n = 1)	Client (n = 4); provider (n = 3); organizational (n = 2)	Provider (n = 4); organizational (n = 2)	Provider (n = 3); organizational (n = 3)
Range of sample sizes	10 (providers, organizational, policy) to 86 (providers)	5 (providers) to 300 (providers)	10 (organizational) to 1650 (clients)	10 (organizational) to 200 (clients)	5 (providers) to 163 (providers)	20 (organizational) to 124 (providers)
Countries	Nepal, Ethiopia, Nigeria, Vietnam, India, South Africa, Kenya, Mozambique, Zimbabwe	Rwanda, Myanmar (Burma), Thailand, Kenya, Mozambique	Chile, Nepal, South Africa, Mozambique, Ethiopia	Iraq, Myanmar (Burma), Thailand, Ukraine, Rwanda, Sierra Leone, South Africa	Brazil, Ukraine, Kenya, Vietnam	Nepal, Ethiopia, Nigeria
Adaption approaches	Translated/back-translated (n = 4); stakeholder feedback (n = 4); qualitative study (n = 3); expert feedback (n = 4)	Translated/back-translated (n = 4); stakeholder feedback (n = 5); expert feedback (n = 1)	Translated/back-translated (n = 5); stakeholder feedback (n = 2); cognitive interviewing (n = 1)	Translated/back-translated (n = 4); stakeholder feedback (n = 3); qualitative study (n = 1); expert feedback (n = 1)	Translated/back-translated (n = 3); stakeholder feedback (n = 1); expert feedback (n = 2)	Translated/back-translated (n = 3); stakeholder feedback (n = 1); expert feedback (n = 1); cognitive interviewing (n = 1)
Validation approaches	Pilot-tested, results corroborated with field notes (n = 1)	n = 0	n = 0	n = 0	n = 0	n = 0
Comments	Challenge choosing constructs for adaptation; need ways to translate scales into qualitative questions	Limited response variability; positive response bias; long; some questions not relevant for providers or clients	Straightforward and short; easy to translate; unsure if valid	Some questions not relevant for clients or providers; too many hypotheticals; limited response variability; positive response bias; long	Abstract	Concept not easy to assess using structured questionnaires; prefer qualitative interviews

Note: Measures reported as used by only one investigator, or used only in a high-income country setting, are not included in Table 4. Responses related to the Acceptability of Intervention, Intervention Appropriateness, and Feasibility of Intervention Measures were collapsed across the scales as there was complete overlap within respondents for these measures. Responses related to the Applied Mental Health Research implementation measures, which include client-, provider-, organizational- and policy-level scales for several implementation outcomes and contextual determinants, were collapsed for the same reason.
AIM, Acceptability of Intervention Measure; AMHR/mhIST, Applied Mental Health Research/Mental Health Implementation Science Tool; CFIR, Consolidated Framework for Implementation Research; EBPAS, Evidence-Based Practice Attitude Scale; FIM, Feasibility of Intervention Measure; IAM, Appropriateness of Intervention Measure; ORIC, Organization Readiness for Implementing Change; PSAT, Program Sustainability Assessment Tool.

Investigator survey

We invited 107 investigators to participate in the survey or suggest other investigators for participation. Sixty-two investigators responded. We sent survey links to 45 investigators who indicated interest in participation. Thirty-eight investigators started the survey. Table 3 presents the characteristics of the 28 investigators who completed the survey. The majority (61%) were based in the United States, most (82%) were at universities or other academic institutions and almost all (96%) were focused on research as opposed to clinical service delivery or program implementation. Investigators had been involved in a mean of 2.2 implementation studies related to mental health.

Table 4 describes the usage of implementation measures reported by at least two investigators in LMIC settings. The most used implementation measures included the Consolidated Framework for Implementation Research Inner Setting measures (n = 7) (Fernandez et al., 2018), the Program Assessment Sustainability Tool (n = 5) (Luke et al., 2014) and the Acceptability of Intervention Measure, Intervention Appropriateness Measure and Feasibility of Intervention Measure (n = 5) (Weiner et al., 2017). Measures were most commonly used prior to intervention implementation (n = 18) or mid-implementation (n = 18) as opposed to post-implementation (n = 7) and were most often used to assess contextual determinants of implementation effectiveness (n = 20) rather than to assess implementation outcomes (n = 9). Providers were the most common group sampled (n = 25), followed by clients (n = 9). Measures were used in a diverse range of contexts across Latin America, Sub-Saharan Africa, Eastern Europe and South/Southeast Asia. Adaptation approaches were generally limited to translation and back-translation (n = 23) and stakeholder feedback (n = 16), and only one investigator reported conducting any measure validation prior to use (pilot testing). Limited response variability, positive response bias, measure length and item relevance were the most common challenges reported.

Other measures reported as used by individual investigators included the Implementation Leadership Scale (Aarons et al., 2014), the Theory of Planned Behavior measures (Ajzen, 2011), the Feelings Thermometer (ALWIN, 1997), the Systems Usability

Scale (Lewis, 2018), the Organizational Social Context scale (Glisson et al., 2008), several intervention-specific fidelity scales and several measures developed new for individual studies.

Discussion

This study sought to improve quantitative implementation measurement in the field of global mental health by generating consensus recommendations on best practices for measure choice and validation and by surveying the field to understand current practice. Our expert panel concluded that pragmatic concerns are key to choosing between measures and validation approaches. They noted that many quantitative implementation measures are lengthy and identified a trade-off between resources and rigor in the various approaches available for adapting and validating implementation measures in diverse global settings. However, they concluded that in many cases, it is sufficient for investigators to establish the face validity of an implementation measure in a new setting through some combination of reviewing the use of that measure in a similar setting, convening an informal expert and stakeholder panel, conducting translation and back-translation and piloting the measure to confirm its dimensionality and internal reliability. Though confirming the predictive validity of a measure by correlating it with subsequent real-world outcomes would be the gold standard for measure validation, panel members felt this was unnecessary prior to using most implementation measures. Survey results suggested that though several implementation measures have been used or are in use in global mental health studies across a variety of levels and study phases, almost none have been formally validated as part of those studies.

Quantitative measures must be reliable, valid and practical to be useful for implementation research or practice, though comprehensive reviews of published implementation measures have noted that the field faces several major issues. These include the poor distribution of quantitative measures across implementation constructs and analytic levels; a lack of measures with strong psychometric qualities; measure synonymy (the same measure items are sometimes used to measure different constructs), homonymy (different measure items are used to measure the same construct) and instability (measure items are often changed with each use) and the reality that many implementation measures exhibit poor pragmatic qualities (Lewis et al., 2018). Nevertheless, a growing number of strong implementation measures do exist: the challenge for investigators in diverse global settings in choosing and adapting these – or developing new ones – and ensuring that they perform well. Notably, the Psychometric and Pragmatic Evidence Rating Scale has been developed through stakeholder consensus to provide clear criteria for measure quality, both to inform measure development and measure choice (Stanick et al., 2019). In addition, domain-specific resources are increasingly available to support investigators in choosing between manifest and latent indicators of implementation process and outcomes, including the HIV Implementation Outcomes Crosswalk (Li et al., 2020).

Several key limitations should be noted. Our expert panel consisted of only seven members, reflecting the relatively small number of individuals with intersecting expertise in global mental health, implementation science and psychometrics. In response, we opted for depth over breadth and sought to reach panel consensus across a wide range of issues related to measure use and validation, rather than for one or two key questions. Our Delphi panel size is considered acceptable for non-statistical analysis (Rowe and Wright, 1999). All panel procedures were carried out during the first 6 months of the COVID-19 pandemic, meaning procedures were remote and sometimes asynchronous. For our survey, we sampled investigators from NIH RePORTER and Grand Challenges Canada; these are two of the most prolific funders of global mental health implementation research, though this approach likely biased our sample toward investigators based in North America. To mitigate this risk, we used snowball sampling to attempt to identify and recruit other investigators that would have been missed with this approach. Our overall response rate was low, which again may reflect the small number of individuals actively using quantitative measures in their global mental health implementation studies; many investigators we contacted declined to participate because they were not using quantitative implementation measures.

Despite these limitations, our findings may directly support the growing field of global mental health implementation research. We have used our results to compile a set of guidance documents for investigators planning to quantitatively measure latent implementation processes and outcomes in diverse global settings. These include a compendium of available measures across implementation constructs and detailed descriptions of common adaptation and validation approaches. This guidance should facilitate rigorous and replicable implementation research in an area of high need, though it is not intended to be prescriptive, and local investigators are encouraged to adapt and apply the guidance only where it is useful. Moving forward, as the quantity and quality of implementation measures designed for use in for diverse global contexts increase (Aldridge et al., 2022), the standards for measure adaptation and validation may also shift. Less emphasis may be placed on establishing measure validity for the sake of scientific rigor, with a corresponding increased emphasis on measure pragmatic qualities and capacity to inform real-world health service delivery.

References

Aarons GA, Ehrhart MG and Farahnak LR (2014) The implementation leadership scale (ILS): Development of a brief measure of unit level implementation leadership. *Implementation Science* 9, 45. https://doi.org/10.1186/1748-5908-9-45.

Ajzen I (2011) The theory of planned behaviour: Reactions and reflections. *Psychology & Health* 26, 1113–1127. https://doi.org/10.1080/08870446.2011.613995.

Aldridge LR, Kemp CG, Bass JK, Danforth K, Kane JC, Hamdani SU, Marsch LA, Uribe-Restrepo JM, Nguyen AJ and Bolton PA (2022) Psychometric performance of the mental health implementation science tools (mhIST) across six low-and middle-income countries. *Implementation Science Communications* 3, 54

Alwin DF (1997) Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research* 25, 318–340. https://doi.org/10.1177/0049124197025003003.

Avella JR (2016) Delphi panels: Research design, procedures, advantages, and challenges. *International Journal of Doctoral Studies* 11, 305.

Betancourt TS and Chambers DA (2016) Optimizing an era of global mental health implementation science. *JAMA Psychiatry* 73, 99–100. https://doi.org/10.1001/jamapsychiatry.2015.2705.

Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR and Young SL (2018) Best practices for developing and validating scales for health, social, and Behavioral research: A primer. *Frontiers in Public Health* 6(149), 1–18.

Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA and Lowery JC (2009) Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science* 4, 50. https://doi.org/10.1186/1748-5908-4-50.

Eccles MP and Mittman BS (2006) Welcome to implementation science. *Implementation Science* 1(1), 1–3.

Fernandez ME, Walker TJ, Weiner BJ, Calo WA, Liang S, Risendal B, Friedman DB, Tu SP, Williams RS, Jacobs S, Herrmann AK and Kegler MC (2018) Developing measures to assess constructs from the inner setting domain of the consolidated framework for implementation research. *Implementation ScienceS* 13, 52. https://doi.org/10.1186/s13012-018-0736-7.

Glisson C, Landsverk J, Schoenwald S, Kelleher K, Hoagwood KE, Mayberg S and Green P (2008) Assessing the organizational social context (OSC) of mental health services: Implications for research and practice. *Administration and Policy in Mental Health* 35, 98–113. https://doi.org/10.1007/s10488-007-0148-5.

Hull L, Boulton R, Jones F, Boaz A and Sevdalis N (2022) Defining, conceptualizing and evaluating pragmatic qualities of quantitative instruments measuring implementation determinants and outcomes: A scoping and critical review of the literature and recommendations for future research. *Translational Behavioral Medicine* 12, 1049–1064. https://doi.org/10.1093/tbm/ibac064.

Khadjesari Z, Boufkhed S, Vitoratou S, Schatte L, Ziemann A, Daskalopoulou C, Uglik-Marucha E, Sevdalis N and Hull L (2020) Implementation outcome instruments for use in physical healthcare settings: A systematic review. *Implementation Science* 15, 66. https://doi.org/10.1186/s13012-020-01027-6.

Lancet Global Mental Health Group, Chisholm D, Flisher AJ, Lund C, Patel V, Saxena S, Thornicroft G and Tomlinson M (2007) Scale up services for mental disorders: A call for action. *The Lancet* 370, 1241–1252. https://doi.org/10.1016/S0140-6736(07)61242-2.

Lewis JR (2018) The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction* 34, 577–90.

Lewis CC, Fischer S, Weiner BJ, Stanick C, Kim M and Martinez RG (2015) Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science* 10, 155. https://doi.org/10.1186/s13012-015-0342-x.

Lewis CC, Mettert KD, Dorsey CN, Martinez RG, Weiner BJ, Nolen E, Stanick C, Halko H and Powell BJ (2018) An updated protocol for a systematic review of implementation-related measures. *Systematic Reviews* 7, 66. https://doi.org/10.1186/s13643-018-0728-3.

Lewis CC, Weiner BJ, Stanick C and Fischer SM (2015) Advancing implementation science through measure development and evaluation: A study protocol. *Implementation Science* 10, 102.

Li D, Audet C and Schwartz S (2020) HIV implementation outcomes operationalization guide – ISC3I. Available at https://isc3i.isgmh.northwestern.edu/hivoutcomes/ (accessed 5 February 2022).

Linstone HA and Turoff M (1975) *The Delphi Method: Techniques and Applications*. Boston: Addison Wesley Publishing Company.

Luke DA, Calhoun A, Robichaux CB, Elliott MB and Moreland-Russell S (2014) Peer reviewed: The program sustainability assessment tool: A new instrument for public health programs. *Preventing Chronic Disease* 11.

Means AR, Kemp CG, Gwayi-Chore M-C, Gimbel S, Soi C, Sherr K, Wagenaar BH, Wasserheit JN and Weiner BJ (2020) Evaluating and optimizing the consolidated framework for implementation research (CFIR) for use in low- and middle-income countries: A systematic review. *Implementation Science* 15, 17. https://doi.org/10.1186/s13012-020-0977-0.

Mettert K, Lewis C, Dorsey C, Halko H and Weiner B (2020) Measuring implementation outcomes: An updated systematic review of measures' psychometric properties. *Implementation Research and Practice* 1, 2633489520936644. https://doi.org/10.1177/2633489520936644.

Palinkas LA (2014) Qualitative and mixed methods in mental health services and implementation research. *Journal of Clinical Child & Adolescent Psychology* 53(43), 851–61. https://doi.org/10.1080/15374416.2014.910791.

Palinkas LA, Aarons GA, Horwitz S, Chamberlain P, Hurlburt M and Landsverk J (2011) Mixed method designs in implementation research. *Administration and Policy in Mental Health and Mental Health Services Research* 38, 44–53. https://doi.org/10.1007/s10488-010-0314-z.

Patel V, Chisholm D, Dua T, Laxminarayan R and Medina-Mora ME (eds.) (2016) *Mental, Neurological, and Substance Use Disorders: Disease Control Priorities*, 3rd Edn. (Volume 4). Washington, DC: The International Bank for Reconstruction and Development/The World Bank.

Pathare S, Brazinova A and Levav I (2018) Care gap: A comprehensive measure to quantify unmet needs in mental health. *Epidemiology and Psychiatric Sciences* 27, 463–467. https://doi.org/10.1017/S2045796018000100.

Powell BJ, Stanick CF, Halko HM, Dorsey CN, Weiner BJ, Barwick MA, Damschroder LJ, Wensing M, Wolfenden L and Lewis CC (2017) Toward criteria for pragmatic measurement in implementation research and practice: A stakeholder-driven approach using concept mapping. *Implementation Science* 12, 118. https://doi.org/10.1186/s13012-017-0649-x.

Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, Griffey R and Hensley M (2011) Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health* 38, 65–76. https://doi.org/10.1007/s10488-010-0319-7.

Rowe G and Wright G (1999) The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting* 15, 353–75. https://doi.org/10.1016/S0169-2070(99)00018-7.

Stanick CF, Halko HM, Nolen EA, Powell BJ, Dorsey CN, Mettert KD, Weiner BJ, Barwick M, Wolfenden L, Damschroder LJ and Lewis CC (2019) Pragmatic measures for implementation research: Development of the psychometric and pragmatic evidence rating scale (PAPERS). *Translational Behavioral Medicine* 11, 11–20. https://doi.org/10.1093/tbm/ibz164.

Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, Al-Hamzawi A, Alonso J, Andrade L and Borges G (2017) Undertreatment of people with major depressive disorder in 21 countries. *The British Journal of Psychiatry* 210, 119–24.

Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, Abbasi-Kangevari M, Abbastabar H, Abd-Allah F, Abdelalim A, Abdollahi M, Abdollahpour I, Abolhassani H, Aboyans V, Abrams EM, Abreu LG, Abrigo MRM, Abu-Raddad LJ, Abushouk AI, Acebedo A, Ackerman IN, Adabi M, Adamu AA, Adebayo OM, Adekanmbi V, Adelson JD, Adetokunboh OO, Adham D, Afshari M, Afshin A, Agardh EE, Agarwal G, Agesa KM, Aghaali M, Aghamir SMK, Agrawal A, Ahmad T, Ahmadi A, Ahmadi M, Ahmadieh H, Ahmadpour E, Akalu TY, Akinyemi RO, Akinyemiju T, Akombi B, Al-Aly Z, Alam K, Alam N, Alam S, Alam T, Alanzi TM, Albertson SB, Alcalde-Rabanal JE, Alema NM, Ali M, Ali S, Alicandro G, Alijanzadeh M, Alinia C, Alipour V, Aljunid SM, Alla F, Allebeck P, Almasi-Hashiani A, Alonso J, Al-Raddadi RM, Altirkawi KA, Alvis-Guzman N, Alvis-Zakzuk NJ, Amini S, Amini-Rarani M, Aminorroaya A, Amiri F, Amit AML, Amugsi DA, Amul GGH, Anderlini D, Andrei CL, Andrei T, Anjomshoa M, Ansari F, Ansari I, Ansari-Moghaddam

A, Antonio CAT, Antony CM, Antriyandarti E, Anvari D, Anwer R, Arabloo J, Arab-Zozani M, Aravkin AY, Ariani F, Ärnlöv J, Aryal KK, Arzani A, Asadi-Aliabadi M, Asadi-Pooya AA, Asghari B, Ashbaugh C, Atnafu DD, Atre SR, Ausloos F, Ausloos M, Ayala Quintanilla BP, Ayano G, Ayanore MA, Aynalem YA, Azari S, Azarian G, Azene ZN, Babaee E, Badawi A, Bagherzadeh M, Bakhshaei MH, Bakhtiari A, Balakrishnan S, Balalla S, Balassyano S, Banach M, Banik PC, Bannick MS, Bante AB, Baraki AG, Barboza MA, Barker-Collo SL, Barthelemy CM, Barua L, Barzegar A, Basu S, Baune BT, Bayati M, Bazmandegan G, Bedi N, Beghi E, Béjot Y, Bello AK, Bender RG, Bennett DA, Bennitt FB, Bensenor IM, Benziger CP, Berhe K, Bernabe E, Bertolacci GJ, Bhageerathy R, Bhala N, Bhandari D, Bhardwaj P, Bhattacharyya K, Bhutta ZA, Bibi S, Biehl MH, Bikbov B, Bin Sayeed MS, Biondi A, Birihane BM, Bisanzio D, Bisignano C, Biswas RK, Bohlouli S, Bohluli M, Bolla SRR, Boloor A, Boon-Dooley AS, Borges G, Borzì AM, Bourne R, Brady OJ, Brauer M, Brayne C, Breitborde NJK, Brenner H, Briant PS, Briggs AM, Briko NI, Britton GB, Bryazka D, Buchbinder R, Bumgarner BR, Busse R, Butt ZA, Caetano dos Santos FL, Cámera LLA, Campos-Nonato IR, Car J, Cárdenas R, Carreras G, Carrero JJ, Carvalho F, Castaldelli-Maia JM, Castañeda-Orjuela CA, Castelpietra G, Castle CD, Castro F, Catalá-López F, Causey K, Cederroth CR, Cercy KM, Cerin E, Chandan JS, Chang AR, Charlson FJ, Chattu VK, Chaturvedi S, Chimed-Ochir O, Chin KL, Cho DY, Christensen H, Chu D-T, Chung MT, Cicuttini FM, Ciobanu LG, Cirillo M, Collins EL, Compton K, Conti S, Cortesi PA, Costa VM, Cousin E, Cowden RG, Cowie BC, Cromwell EA, Cross DH, Crowe CS, Cruz JA, Cunningham M, Dahlawi SMA, Damiani G, Dandona L, Dandona R, Darwesh AM, Daryani A, Das JK, Das Gupta R, das Neves J, Dávila-Cervantes CA, Davletov K, De Leo D, Dean FE, DeCleene NK, Deen A, Degenhardt L, Dellavalle RP, Demeke FM, Demsie DG, Denova-Gutiérrez E, Dereje ND, Dervenis N, Desai R, Desalew A, Dessie GA, Dharmaratne SD, Dhungana GP, Dianatinasab M, Diaz D, Dibaji Forooshani ZS, Dingels ZV, Dirac MA, Djalalinia S, Do HT, Dokova K, Dorostkar F, Doshi CP, Doshmangir L, Douiri A, Doxey MC, Driscoll TR, Dunachie SJ, Duncan BB, Duraes AR, Eagan AW, Ebrahimi Kalan M, Edvardsson D, Ehrlich JR, El Nahas N, El Sayed I, El Tantawi M, Elbarazi I, Elgendy IY, Elhabashy HR, El-Jaafary SI, Elyazar IR, Emamian MH, Emmons-Bell S, Erskine HE, Eshrati B, Eskandarieh S, Esmaeilnejad S, Esmaeilzadeh F, Esteghamati A, Estep K, Etemadi A, Etisso AE, Farahmand M, Faraj A, Fareed M, Faridnia R, Farinha CSS, Farioli A, Faro A, Faruque M, Farzadfar F, Fattahi N, Fazlzadeh M, Feigin VL, Feldman R, Fereshtehnejad S-M, Fernandes E, Ferrari AJ, Ferreira ML, Filip I, Fischer F, Fisher JL, Fitzgerald R, Flohr C, Flor LS, Foigt NA, Folayan MO, Force LM, Fornari C, Foroutan M, Fox JT, Freitas M, Fu W, Fukumoto T, Furtado JM, Gad MM, Gakidou E, Galles NC, Gallus S, Gamkrelidze A, Garcia-Basteiro AL, Gardner WM, Geberemariyam BS, Gebrehiwot AM, Gebremedhin KB, Gebreslassie AAAA, Gershberg Hayoon A, Gething PW, Ghadimi M, Ghadiri K, Ghafourifard M, Ghajar A, Ghamari F, Ghashghaee A, Ghiasvand H, Ghith N, Gholamian A, Gilani SA, Gill PS, Gitimoghaddam M, Giussani G, Goli S, Gomez RS, Gopalani SV, Gorini G, Gorman TM, Gottlich HC, Goudarzi H, Goulart AC, Goulart BNG, Grada A, Grivna M, Grosso G, Gubari MIM, Gugnani HC, Guimaraes ALS, Guimarães RA, Guled RA, Guo G, Guo Y, Gupta R, Haagsma JA, Haddock B, Hafezi-Nejad N, Hafiz A, Hagins H, Haile LM, Hall BJ, Halvaei I, Hamadeh RR, Hamagharib Abdullah K, Hamilton EB, Han C, Han H, Hankey GJ, Haro JM, Harvey JD, Hasaballah AI, Hasanzadeh A, Hashemian M, Hassanipour S, Hassankhani H, Havmoeller RJ, Hay RJ, Hay SI, Hayat K, Heidari B, Heidari G, Heidari-Soureshjani R, Hendrie D, Henrikson HJ, Henry NJ, Herteliu C, Heydarpour F, Hird TR, Hoek HW, Hole MK, Holla R, Hoogar P, Hosgood HD, Hosseinzadeh M, Hostiuc M, Hostiuc S, Househ M, Hoy DG, Hsairi M, Hsieh VC, Hu G, Huda TM, Hugo FN, Huynh CK, Hwang B-F, Iannucci VC, Ibitoye SE, Ikuta KS, Ilesanmi OS, Ilic IM, Ilic MD, Inbaraj LR, Ippolito H, Irvani SSN, Islam MM, Islam M, Islam SMS, Islami F, Iso H, Ivers RQ, Iwu CCD, Iyamu IO, Jaafari J, Jacobsen KH, Jadidi-Niaragh F, Jafari H, Jafarinia M, Jahagirdar D, Jahani MA, Jahanmehr N, Jakovljevic M, Jalali A, Jalilian F, James SL, Janjani H, Janodia MD, Jayatilleke AU, Jeemon P, Jenabi E, Jha RP, Jha V, Ji JS, Jia P, John O, John-Akinola YO, Johnson CO, Johnson SC, Jonas JB, Joo T, Joshi A, Jozwiak JJ, Jürisson M, Kabir A, Kabir Z, Kalani H, Kalani R, Kalankesh LR, Kalhor R, Kamiab Z, Kanchan T, Karami

Matin B, Karch A, Karim MA, Karimi SE, Kassa GM, Kassebaum NJ, Katikireddi SV, Kawakami N, Kayode GA, Keddie SH, Keller C, Kereselidze M, Khafaie MA, Khalid N, Khan M, Khatab K, Khater MM, Khatib MN, Khayamzadeh M, Khodayari MT, Khundkar R, Kianipour N, Kieling C, Kim D, Kim Y-E, Kim YJ, Kimokoti RW, Kisa A, Kisa S, Kissimova-Skarbek K, Kivimäki M, Kneib CJ, Knudsen AKS, Kocarnik JM, Kolola T, Kopec JA, Kosen S, Koul PA, Koyanagi A, Kravchenko MA, Krishan K, Krohn KJ, Kuate Defo B, Kucuk Bicer B, Kumar GA, Kumar M, Kumar P, Kumar V, Kumaresh G, Kurmi OP, Kusuma D, Kyu HH, La Vecchia C, Lacey B, Lal DK, Lalloo R, Lam JO, Lami FH, Landires I, Lang JJ, Lansingh VC, Larson SL, Larsson AO, Lasrado S, Lassi ZS, Lau KM-M, Lavados PM, Lazarus JV, Ledesma JR, Lee PH, Lee SWH, LeGrand KE, Leigh J, Leonardi M, Lescinsky H, Leung J, Levi M, Lewington S, Li S, Lim L-L, Lin C, Lin R-T, Linehan C, Linn S, Liu H-C, Liu S, Liu Z, Looker KJ, Lopez AD, Lopukhov PD, Lorkowski S, Lotufo PA, Lucas TCD, Lugo A, Lunevicius R, Lyons RA, Ma J, MacLachlan JH, Maddison ER, Maddison R, Madotto F, Mahasha PW, Mai HT, Majeed A, Maled V, Maleki S, Malekzadeh R, Malta DC, Mamun AA, Manafi A, Manafi N, Manguerra H, Mansouri B, Mansournia MA, Mantilla Herrera AM, Maravilla JC, Marks A, Martins-Melo FR, Martopullo I, Masoumi SZ, Massano J, Massenburg BB, Mathur MR, Maulik PK, McAlinden C, McGrath JJ, McKee M, Mehndiratta MM, Mehri F, Mehta KM, Meitei WB, Memiah PTN, Mendoza W, Menezes RG, Mengesha EW, Mengesha MB, Mereke A, Meretoja A, Meretoja TJ, Mestrovic T, Miazgowski B, Miazgowski T, Michalek IM, Mihretie KM, Miller TR, Mills EJ, Mirica A, Mirrakhimov EM, Mirzaei H, Mirzaei M, Mirzaei-Alavijeh M, Misganaw AT, Mithra P, Moazen B, Moghadaszadeh M, Mohamadi E, Mohammad DK, Mohammad Y, Mohammad Gholi Mezerji N, Mohammadian-Hafshejani A, Mohammadifard N, Mohammadpourhodki R, Mohammed S, Mokdad AH, Molokhia M, Momen NC, Monasta L, Mondello S, Mooney MD, Moosazadeh M, Moradi G, Moradi M, Moradi-Lakeh M, Moradzadeh R, Moraga P, Morales L, Morawska L, Moreno Velásquez I, Morgado-da-Costa J, Morrison SD, Mosser JF, Mouodi S, Mousavi SM, Mousavi Khaneghah A, Mueller UO, Munro SB, Muriithi MK, Musa KI, Muthupandian S, Naderi M, Nagarajan AJ, Nagel G, Naghshtabrizi B, Nair S, Nandi AK, Nangia V, Nansseu JR, Nayak VC, Nazari J, Negoi I, Negoi RI, Netsere HBN, Ngunjiri JW, Nguyen CT, Nguyen J, Nguyen M, Nguyen M, Nichols E, Nigatu D, Nigatu YT, Nikbakhsh R, Nixon MR, Nnaji CA, Nomura S, Norrving B, Noubiap JJ, Nowak C, Nunez-Samudio V, Oţoiu A, Oancea B, Odell CM, Ogbo FA, Oh I-H, Okunga EW, Oladnabi M, Olagunju AT, Olusanya BO, Olusanya JO, Oluwasanu MM, Omar Bali A, Omer MO, Ong KL, Onwujekwe OE, Orji AU, Orpana HM, Ortiz A, Ostroff SM, Otstavnov N, Otstavnov SS, Øverland S, Owolabi MO, P a M, Padubidri JR, Pakhare AP, Palladino R, Pana A, Panda-Jonas S, Pandey A, Park E-K, Parmar PGK, Pasupula DK, Patel SK, Paternina-Caicedo AJ, Pathak A, Pathak M, Patten SB, Patton GC, Paudel D, Pazoki Toroudi H, Peden AE, Pennini A, Pepito VCF, Peprah EK, Pereira A, Pereira DM, Perico N, Pham HQ, Phillips MR, Pigott DM, Pilgrim T, Pilz TM, Pirsaheb M, Plana-Ripoll O, Plass D, Pokhrel KN, Polibin RV, Polinder S, Polkinghorne KR, Postma MJ, Pourjafar H, Pourmalek F, Pourmirza Kalhori R, Pourshams A, Poznańska A, Prada SI, Prakash V, Pribadi DRA, Pupillo E, Quazi Syed Z, Rabiee M, Rabiee N, Radfar A, Rafiee A, Rafiei A, Raggi A, Rahimi-Movaghar A, Rahman MA, Rajabpour-Sanati A, Rajati F, Ramezanzadeh K, Ranabhat CL, Rao PC, Rao SJ, Rasella D, Rastogi P, Rathi P, Rawaf DL, Rawaf S, Rawal L, Razo C, Redford SB, Reiner RC, Reinig N, Reitsma MB, Remuzzi G, Renjith V, Renzaho AMN, Resnikoff S, Rezaei N, Rezai M sadegh, Rezapour A, Rhinehart P-A, Riahi SM, Ribeiro ALP, Ribeiro DC, Ribeiro D, Rickard J, Roberts NLS, Roberts S, Robinson SR, Roever L, Rolfe S, Ronfani L, Roshandel G, Roth GA, Rubagotti E, Rumisha SF, Sabour S, Sachdev PS, Saddik B, Sadeghi E, Sadeghi M, Saeidi S, Safi S, Safiri S, Sagar R, Sahebkar A, Sahraian MA, Sajadi SM, Salahshoor MR, Salamati P, Salehi Zahabi S, Salem H, Salem MRR, Salimzadeh H, Salomon JA, Salz I, Samad Z, Samy AM, Sanabria J, Santomauro DF, Santos IS, Santos JV, Santric-Milicevic MM, Saraswathy SYI, Sarmiento-Suárez R, Sarrafzadegan N, Sartorius B, Sarveazad A, Sathian B, Sathish T, Sattin D, Sbarra AN, Schaeffer LE, Schiavolin S, Schmidt MI, Schutte AE, Schwebel DC, Schwendicke F, Senbeta AM, Senthilkumaran S, Sepanlou SG, Shackelford KA, Shadid J, Shahabi S, Shaheen AA, Shaikh MA, Shalash AS,

Shams-Beyranvand M, Shamsizadeh M, Shannawaz M, Sharafi K, Sharara F, Sheena BS, Sheikhtaheri A, Shetty RS, Shibuya K, Shiferaw WS, Shigematsu M, Shin JI, Shiri R, Shirkoohi R, Shrime MG, Shuval K, Siabani S, Sigfusdottir ID, Sigurvinsdottir R, Silva JP, Simpson KE, Singh A, Singh JA, Skiadaresi E, Skou STS, Skryabin VY, Sobngwi E, Sokhan A, Soltani S, Sorensen RJD, Soriano JB, Sorrie MB, Soyiri IN, Sreeramareddy CT, Stanaway JD, Stark BA, Ştefan SC, Stein C, Steiner C, Steiner TJ, Stokes MA, Stovner LJ, Stubbs JL, Sudaryanto A, Sufiyan MB, Sulo G, Sultan I, Sykes BL, Sylte DO, Szócska M, Tabarés-Seisdedos R, Tabb KM, Tadakamadla SK, Taherkhani A, Tajdini M, Takahashi K, Taveira N, Teagle WL, Teame H, Tehrani-Banihashemi A, Teklehaimanot BF, Terrason S, Tessema ZT, Thankappan KR, Thomson AM, Tohidinik HR, Tonelli M, Topor-Madry R, Torre AE, Touvier M, Tovani-Palone MRR, Tran BX, Travillian R, Troeger CE, Truelsen TC, Tsai AC, Tsatsakis A, Tudor Car L, Tyrovolas S, Uddin R, Ullah S, Undurraga EA, Unnikrishnan B, Vacante M, Vakilian A, Valdez PR, Varughese S, Vasankari TJ, Vasseghian Y, Venketasubramanian N, Violante FS, Vlassov V, Vollset SE, Vongpradith A, Vukovic A, Vukovic R, Waheed Y, Walters MK, Wang J, Wang Y, Wang Y-P, Ward JL, Watson A, Wei J, Weintraub RG, Weiss DJ, Weiss J, Westerman R, Whisnant JL, Whiteford HA, Wiangkham T, Wiens KE, Wijeratne T, Wilner LB, Wilson S, Wojtyniak B, Wolfe CDA, Wool EE, Wu A-M, Wulf Hanson S, Wunrow HY, Xu G, Xu R, Yadgir S, Yahyazadeh Jabbari SH, Yamagishi K, Yaminfirooz M, Yano Y, Yaya S, Yazdi-Feyzabadi V, Yearwood JA, Yeheyis TY, Yeshitila YG, Yip P, Yonemoto N, Yoon S-J, Yoosefi Lebni J, Younis MZ, Younker TP, Yousefi Z, Yousefifard M, Yousefinezhadi T, Yousuf AY, Yu C, Yusefzadeh H, Zahirian Moghadam T, Zaki L, Zaman SB, Zamani M, Zamanian M, Zandian H, Zangeneh A, Zastrozhin MS, Zewdie KA, Zhang Y, Zhang Z-J, Zhao JT, Zhao Y, Zheng P, Zhou M, Ziapour A, Zimsen SRM, Naghavi M and Murray CJL (2020) Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet* **396**, 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9.

Wagenaar BH, Hammett WH, Jackson C, Atkins DL, Belus JM and Kemp CG (2020) Implementation outcomes and strategies for depression interventions in low- and middle-income countries: A systematic review. *Global Mental Health* **7**, e7. https://doi.org/10.1017/gmh.2020.1.

Waltz TJ, Powell BJ, Fernández ME, Abadie B and Damschroder LJ (2019) Choosing implementation strategies to address contextual barriers: Diversity in recommendations and future directions. *Implementation Science* **14**, 42. https://doi.org/10.1186/s13012-019-0892-4

Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, Boynton MH and Halko H (2017) Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science* **12**, 108.

Willmeroth T, Wesselborg B and Kuske S (2019) Implementation outcomes and indicators as a new challenge in health services research: A systematic scoping review. *Inquiry* **56**, 0046958019861257. https://doi.org/10.1177/0046958019861257