TBM

ORIGINAL RESEARCH

Pragmatic measures for implementation research: development of the Psychometric and Pragmatic Evidence Rating Scale (PAPERS)

Cameo F. Stanick,[1] Heather M. Halko,[2] Elspeth A. Nolen,[3] Byron J. Powell, [4] Caitlin N. Dorsey,[5] Kayne D. Mettert,[5] Bryan J. Weiner,[6] Melanie Barwick,[7] Luke Wolfenden,[8] Laura J. Damschroder,[9] Cara C. Lewis[5].

[1]Hathaway-Sycamores Child and Family Services, Pasadena, CA, USA

[2]University of Montana, Missoula, MT, USA

[3]Department of Global Health, University of Washington, Seattle, WA, USA

[4]Brown School, Washington University in St. Louis, St. Louis, MO, USA

[5]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA

[6]Departments of Global Health and Health Services, University of Washington, Seattle, WA, USA

[7]Hospital for Sick Children and University of Toronto, Toronto, Canada

[8]School of Medicine and Public Health, The University of Newcastle, Newcastle, Australia

[9]VA Ann Arbor Healthcare System, Center for Clinical Management Research (CCMR), Ann Arbor, MI, USA

Correspondence to: C F Stanick, cstanick@hscfs.org

Cite this as: *TBM* 2021;11:11–20
doi: 10.1093/tbm/ibz164

Abstract

The use of reliable, valid measures in implementation practice will remain limited without pragmatic measures. Previous research identified the need for pragmatic measures, though the characteristic identification used only expert opinion and literature review. Our team completed four studies to develop a stakeholder-driven pragmatic rating criteria for implementation measures. We published Studies 1 (identifying dimensions of the pragmatic construct) and 2 (clarifying the internal structure) that engaged stakeholders—participants in mental health provider and implementation settings—to identify 17 terms/phrases across four categories: Useful, Compatible, Acceptable, and Easy. This paper presents Studies 3 and 4: a Delphi to ascertain stakeholder-prioritized dimensions within a mental health context, and a pilot study applying the rating criteria. Stakeholders ($N = 26$) participated in a Delphi and rated the relevance of 17 terms/phrases to the pragmatic construct. The investigator team further defined and shortened the list, which were piloted with 60 implementation measures. The Delphi confirmed the importance of all pragmatic criteria, but provided little guidance on relative importance. The investigators removed or combined terms/phrases to obtain 11 criteria. The 6-point rating system assigned to each criterion demonstrated sufficient variability across items. The grey literature did not add critical information. This work produced the first stakeholder-driven rating criteria to assess whether measures are pragmatic. The Psychometric and Pragmatic Evidence Rating Scale (PAPERS) combines the pragmatic criteria with psychometric rating criteria, from previous work. Use of PAPERS can inform development of implementation measures and to assess the quality of existing measures.

Key words

Implementation science, Measurement, Psychometric, Pragmatic measure

Implications

Practice: The development of user-centered pragmatic rating criteria will help practitioners and implementation intermediaries identify psychometrically and pragmatically strong measures to plan, monitor, and evaluate their implementation efforts.

Policy: Policymakers who want to monitor policy implementation should encourage use of implementation measures that are psychometrically strong and pragmatic.

Research: Future research should focus on using the Psychometric and Pragmatic Evidence Rating Scale criteria to (a) evaluate existing implementation measures and (b) develop new measurement tools that align with stakeholder needs and priorities.

BACKGROUND

A significant contribution of implementation science to "real world" implementation practice would be the development of reliable and valid measures for assessing implementation context, processes, and outcomes. With such measures in hand, practitioners could assess local barriers to implementation and select implementation strategies that address those barriers, monitor implementation progress making mid-course corrections as needed, and evaluate implementation success as well as the factors that facilitate or inhibit it. However, as Glasgow and Riley stated in their call to action for pragmatic clinical outcome measures, practitioners are unlikely to utilize measures, even psychometrically strong ones, if they are not also "pragmatic" [1]. Indeed, stakeholders in dissemination and implementation initiatives, including practitioners, intermediaries, and other knowledge users in relation to implementation work, may not be trained in the use of standardized, quantitative implementation measures. Similar to issues surrounding pragmatic qualities of clinical outcome measures, training requirements for using implementation measures may be unclear, require specialized education, be too lengthy, or have a time burden to administer, score, and interpret that make their use unrealistic in a practice setting. These pragmatic measure properties identified by Glasgow and Riley offer important considerations to measure developers, but surprisingly, to the best of our knowledge, stakeholders had not been queried about the features that make measures more or less practical from their perspective. If this knowledge gap is not addressed, we cannot ensure measures are truly pragmatic because implementation stakeholders are the ultimate judge—they will use or reject measures based on their perception of whether a measure is pragmatic.

The importance of pragmatic measures is captured in a movement toward developing clinical outcome measures that are both psychometrically and pragmatically strong. For instance, within the mental health field, the Patient Health Questionnaire (PHQ) was developed with pragmatic qualities in mind (e.g., length: 2- and 9-item versions are available; cost: free) and has been established as a psychometrically strong, accessible measure for depression [2]. The PHQ also includes items that are "actionable"—that is, items are keyed to specific intervention strategies found in evidence-based practices (EBPs) for depression based on national norms [2]. There is a similar need to focus on features such as these—development of psychometrically and pragmatically strong measures to support implementation efforts to inform the selection of implementation strategies, offer an assessment of the process and progress of implementation, and generate actionable outcome information. Although some implementation science measures have been psychometrically evaluated [3,4], to date, there has been little explicit focus on developing measures that are pragmatic, and we have no capacity to evaluate the quality of their pragmatic properties. Moreover, stakeholders may prioritize different pragmatic measure properties for clinical outcome measures versus implementation measures.

In order to inform implementation efforts both in and outside of research projects, the overarching objective of our research is to develop a set of pragmatic rating criteria for implementation measures that could be combined with psychometric rating criteria and used for measure development, evaluation, and selection. Consistent with the focus of our funding agency, we primarily engaged implementation stakeholders situated in the mental health field, where implementation science is rapidly advancing. To achieve this objective, we undertook four studies, two of which have been published previously. In Study 1, for which the aim was domain delineation (i.e., ensuring that all constructs that are sufficiently differentiated from similar constructs were identified), we systematically defined the properties of pragmatic measures using both deductive and inductive methods [5]. First, we conducted a literature review which revealed 37 terms and phrases related to the term pragmatic. To supplement our search, we conducted interviews with stakeholders (N = 7) representing diverse mental health settings in which implementation had occurred (i.e., school-based mental health setting, community mental health clinics, international implementation intermediary organization, residential treatment center, hospital, state-level mental health policy office), and discovered 10 additional terms and phrases related to the pragmatic construct. After combining results across methods and removing redundancies, we had a final list of 47 unique terms and phrases reflective of pragmatic measures properties. In Study 2, the aim was clarifying the internal structure of the pragmatic construct. We asked an expanded group of stakeholders (N = 24), including those who participated in the first study, to group the 47 terms and phrases into conceptually distinct categories utilizing concept mapping methodology. They were then asked to label each category based on the theme or contents of terms, and then rate the clarity and importance of each term or phrase. Study 2 yielded a four-cluster solution, with the following labels and associated properties: Useful [n = 15 terms/phrases], Compatible [n = 6 terms/phrases], Acceptable [n = 7 terms/phrases], and Easy [n = 19 terms/phrases]) [7].

The present manuscript describes the methods and results of Study 3, a Delphi process for achieving consensus among priority pragmatic properties, and Study 4, a pilot of the established criteria. We also provide our cumulative product that we hope will inform future systematic reviews and measure development studies: a pragmatic measures quality rating scale. The pragmatic rating scale will be combined with the psychometric rating scale from previous published work to form the Psychometric and Pragmatic Evidence Rating Scale (PAPERS) [4,6]. All studies employed user-centered design principles, involving stakeholders to ensure that our conceptual and operational definitions of the pragmatic construct were accurate and relevant to their experiences and contexts.

STUDY 3: DELPHI PROCESS

Method

Participants

We contacted 51 stakeholders representing different professional groups and sectors within the mental health field via snowball sampling, including the stakeholder panel who participated in our earlier work [5,7]. Specifically, implementation contexts and roles represented practitioner organization leadership, including hospitals, and implementation intermediary agency staff. Previous research suggests that a sample size of 12 is sufficient to achieve reliable, efficient consensus on a topic that requires a relatively homogenous (within the same discipline) panel [8]. A total of 26 unique stakeholders responded and agreed to participate in the Delphi process. Fourteen of the total 26 participants completed both of the two rounds. Six participants completed only round 1 and dropped out; therefore, an additional 6 were recruited to replace them to complete the second round. We opted to include a slightly higher than recommended sample size (26, rather than 12 cited above) to allow for more variation of professional context for implementation. Of those participants who completed both rounds, 64% were female, 100% were Caucasian with two international participants (Canada; western European).

Procedure

The results from the concept mapping study required refinement to ensure the Delphi activity would be productive; there were too many terms for stakeholders to consider and differentiate. The investigator team completed a qualitative pruning process to identify which terms and phrases to retain based on data from each preceding study, with the goal to retain those that appeared most important to stakeholders in describing the pragmatic measures construct. Members of the investigative team (C.F.S., H.M.H., C.N.D., B.J.W., B.J.P., and C.C.L.) independently reviewed the results from the concept mapping activity (e.g., category groupings, clarity ratings, and importance ratings) and then made independent decisions about any additional terms and phrases that should be maintained, removed, or combined. During this process, the team also considered alternative wording for terms and phrases that were identified through both the literature review and stakeholder interviews to put forth the most clear and concise verbiage (e.g., “produces reliable and valid results” was chosen over “psychometrically strong or valid”). Then, the independent results were compared, with agreements to cut executed, and disagreements further discussed and final decisions made. The original list of 47 terms and phrases describing the pragmatic measures construct was reduced to 17, while maintaining the four factor solution emergent from the concept mapping study (Acceptable, Compatible, Easy, Useful; see Table 1 for final categories and terms included in the Delphi process).

A modified, multi-round Delphi process, which can be used to transform expert opinion into group consensus [9], was used to further refine the pragmatic measure criteria by exploring those that were considered most important to the majority of stakeholders (i.e., 80% stakeholder agreement). We started with a specified set of terms and phrases to be sorted, rather than an open ended question(s), to help determine what would be sorted [10]. The list of 17 terms and phrases was entered into an online web-based activity and then distributed to stakeholders into multiple rounds spaced 3 weeks apart.

In the first round of the Delphi process, participants were asked to rate the relative importance of each term or phrase. Typically, Delphi rounds include some form of quantitative data collection either through a ranking or rating system of some kind [11]. For this study, rating was done by asking participants to distribute 100 points across the terms/phrases within each of the four categories pertaining to pragmatic measurement. That is, stakeholders assigned more points to the terms and phrases that they believed best exemplified the properties of an Acceptable pragmatic measure. Participants then repeated this task for each remaining category (i.e., Useful, Compatible, and Easy). Finally, stakeholders were asked to redistribute 100 points across all 17 terms and phrases, regardless of category, to reflect the relative importance to the pragmatic measurement construct overall. Utilizing a 100-point rating system in this way prevented participants from suggesting that all terms and phrases were equally highly relevant, which itself would result in a pragmatic rating system that was not pragmatic or feasible. Once the first round was completed, measures of central tendency (i.e., mean, median, and mode) and the interquartile range for scores were calculated and included as data for participants to consider within the second round of the Delphi.

In the second Delphi round, first round participants were invited to repeat the same activity

Table 1 | Final list of terms and phrases for the Delphi

Useful	Compatible
Produces reliable and valid results	Applicable
Informs clinical or organizational decision-making	Fits organizational activities
Acceptable	Easy
Creates a low social desirability bias	Uses accessible language
Relevant	Efficient
Offers relative advantage over existing	Feasible
methods	Easy to interpret
Acceptable (by staff and clients)	Creates low assessor burden (ease of
Low cost	training, scoring, administration time)
	Items not wordy
	Completed with ease
	Brief

to distribute points among the terms and phrases within each category, as well as distribute the same number of points overall. However, in this round, they were also provided with the measures of central tendency and interquartile ranges and asked to use this information when completing their rating tasks. If stakeholders chose to respond outside of the interquartile range, they were asked to provide a qualitative response explaining why their rating differed from the majority opinion. Subsequent rounds (a third, and a fourth if needed) would be undertaken if consensus was not yet achieved.

RESULTS

Stakeholder participant consensus was achieved for all 17 criteria in both rating tasks after only two rounds, suggesting that participants have consensus and share similar opinions about the importance of pragmatic rating criteria. The global rating achieved full consensus (i.e., 80% stakeholder agreement or higher), whereas stakeholders agreed upon 16 of the 17 criteria when asked to rate within categories. One criterion—Easy to interpret—fell below the consensus cut-off (i.e., 5 of 20 stakeholders rated the item outside of the interquartile range); however, this criterion was one of the most highly rated of the criteria and five stakeholder participants fell outside of the interquartile range only because they rated the criterion as being even more important than the interquartile range. Thus, all stakeholders believed that "easy to interpret" was a very important property of pragmatic measures. Table 2 shows the mean scores and standard deviations for all terms and phrases within categories, and those for all terms and phrases.

STUDY 4: PRAGMATIC RATING CRITERIA

Method

Procedure

Although the Delphi activity confirmed that certain properties from the concept mapping are important to stakeholders (e.g., acceptable [by staff and clients]; *creates low assessor burden* [ease of training, scoring, administration time]; see Table 2), the Delphi results did not yield a parsimonious set of properties that would translate into pragmatic measures rating criteria, as we had hoped. There were simply too many terms and phrases rated as important. To address this, the investigative team (C.F.S., H.M.H., C.N.D., B.J.W., B.J.P., and C.C.L.) utilized multiple inputs (e.g., data from the concept mapping and Delphi processes, extant literature, International Advisory Board (IAB; $N = 9$; Melanie Barwick, Laura Damschroder, Jill Francis, Jeremy Grimshaw, Brian Mittman, John Ovretveit, Rob Sanson-Fisher, Michel Wensing, Luke Wolfenden) guidance and refinement activities to determine which properties should be retained, collapsed, or removed. Our IAB members were identified through our professional networks as leading experts in implementation science who have published on measurement issues. We conducted comparison analyses to determine if a term or phrase was identified in the literature and stakeholder interviews [5], prioritized in the concept mapping process [7], and/or rated highly in the Delphi process. Terms or phrases were retained when they met all of these criteria. Additionally, similarly worded terms or phrases were collapsed/combined to represent a single dimension (e.g., "items not wordy" and "brief" were combined to "brief").

Once the final set of properties were identified, each was assigned a six-point numeric rating system consistent with our team's approach to assessing psychometric properties [6]. This would allow measures to be rated from "poor" (−1) to "excellent" (4) with respect to evidence of various pragmatic properties. Our goal was to create descriptive anchors for each number that would allow for reliable application across raters; thus, the investigator team developed the anchors utilizing concept mapping and Delphi results, group discussion regarding increasing dimensions of specific criteria based on their experience with implementation measures, and sample searches to determine where variation breaks occurred (e.g., costs differences in implementation measures from free/public domain to hundreds of dollars per use).

Our final step was to pilot the rating criteria with a sample of implementation measures to inform revisions to our anchors and offer evidence that our rating system produces variable, valid results and rating information. In parallel work funded by the same grant award, our team has conducted systematic reviews of implementation measures that map to the five domains of the Consolidated Framework for Implementation Research (CFIR) [12] and the Implementation Outcomes Framework (IOF) [13]. For the present pilot study, we applied the pragmatic rating criteria to 10 randomly selected measures from those two frameworks from the following six domains (CFIR; IOF; 60 measures total): (a) Outer setting; (b) Inner setting, (c) Characteristics of Individuals, (d) Intervention Characteristics, (e) Process, and (f) Outcomes. Said differently, we piloted our rating criteria on existing implementation measures that crossed a large spectrum of implementation-specific constructs. Our previous work helped us to identify which constructs, and corresponding measures, to select.

The 60 measures we used for piloting our rating criteria and every identified instance of empirical use of those measures in peer-reviewed publications were compiled into packets for piloting the rating criteria. To supplement the peer-reviewed literature that was likely to contain little information regarding pragmatic qualities (as this is not yet common practice), we included a Google search

Table 2 | Descriptive statistics results for both rounds of the Delphi

Round 1	Mean	SD	Round 2	Mean	SD
Useful					
Produces reliable and valid results	51.75	15.92	Produces reliable and valid results	53.10	9.32
Informs clinical or organizational decision-making	48.25	15.92	Informs clinical or organizational decision-making	46.90	9.32
Compatible					
Applicable	42.50	26.28	Applicable	43.50	13.87
Fits organizational activities	57.50	26.28	Fits organizational activities	56.50	13.87
Acceptable					
Creates a low social desirability bias	9.00	8.21	Creates a low social desirability bias	8.65	2.74
Relevant	15.50	9.16	Relevant	20.00	12.14
Offers relative advantage over existing methods	17.25	11.64	Offers relative advantage over existing methods	18.75	7.76
Acceptable (by staff and clients)	39.75	18.88	Acceptable (by staff and clients)	36.35	11.36
Low cost	18.5	12.78	Low cost	16.25	5.60
Easy					
Uses accessible language	12.75	8.50	Uses accessible language	12.65	3.13
Efficient	14.25	11.29	Efficient	13.95	5.13
Feasible	6.75	5.49	Feasible	9.55	4.21
Easy to interpret	13.5	5.91	Easy to interpret	15.10	3.57
Creates low assessor burden (ease of training, scoring, administration time)	18.90	12.29	Creates low assessor burden (ease of training, scoring, administration time)	18.70	7.50
Items not wordy	7.10	5.92	Items not wordy	8.05	3.61
Completed with ease	16.75	7.83	Completed with ease	13.25	5.51
Brief	10.00	5.79	Brief	8.75	2.40
Rate All					
Produces reliable and valid results	9.80	5.43	Produces reliable and valid results	8.75	2.51
Informs clinical or organizational decision- making	9.55	6.41	Informs clinical or organizational decision-making	9.50	1.96
Applicable	5.75	3.46	Applicable	6.30	2.62
Fits organizational activities	6.95	5.53	Fits organizational activities	7.00	2.08
Creates a low social desirability bias	2.40	2.39	Creates a low social desirability bias	3.10	2.10
Relevant	6.30	5.99	Relevant	5.90	1.62
Offers relative advantage over existing methods	3.45	3.14	Offers relative advantage over existing methods	3.85	2.89
Acceptable (by staff and clients)	9.75	4.98	Acceptable (by staff and clients)	9.60	1.85
Low cost	4.40	2.87	Low cost	4.40	0.94
Uses accessible language	4.80	3.30	Uses accessible language	4.85	1.81
Efficient	6.05	6.37	Efficient	5.00	1.52
Feasible	5.20	6.11	Feasible	4.00	1.49
Easy to interpret	5.30	3.70	Easy to interpret	6.50	1.96
Creates low assessor burden (ease of training, scoring, administration time)	7.95	7.45	Creates low assessor burden (ease of training, scoring, administration time)	8.75	2.99
Items not wordy	2.35	2.23	Items not wordy	3.15	2.18
Completed with ease	5.60	3.32	Completed with ease	5.25	1.74
Brief	4.40	3.39	Brief	4.10	1.37

process in our measure review and rating procedures for pragmatic properties. Specifically, for each measure, the measure development article (or the measure's first empirical use if no measure development article was identified) was reviewed by one of three research specialists (E.A.N., H.M.H., and K.D.M.) to identify any information that could be rated using the established pragmatic rating criteria (e.g., a measure was rated as a 3 if it was determined that it cost <US$1 per use; see Table 3). At least two research specialists (to mitigate the influence of algorithmically-personalized search results) then completed a Google search for each measure to identify additional data of relevance to the pragmatic rating criteria. The search for each measure was completed using the following steps, which we based on methods tested in related or other fields wherever possible: (a) the research assistant entered the measure name in quotations in the search field (i.e., "MEASURE NAME") of Google.com (if no formal measure name was provided then the Google search could not be completed); (b) each search was timed to last no longer than 2 min to reflect the amount of time an average

Table 3 | Stakeholder-facing and objective pragmatic rating criteria

Stakeholder Facing Criteria	
Acceptable Category	
Acceptable	
−1	Poor
0	None: The acceptability of the measure was not appraised by staff or clients
1	Minimal/Emerging
2	Adequate
3	Good
4	Excellent
Offers Relative Advantage Over Existing Methods	
−1	Poor
0	None: The relative advantage of the measure over existing methods was not assessed or results were not available
1	Minimal/Emerging
2	Adequate
3	Good
4	Excellent
Easy Category	
Completed with Ease	
−1	Poor
0	None: The ease of completing the measure was not assessed
1	Minimal/Emerging
2	Adequate
3	Good
4	Excellent
Compatible Category	
Appropriate	
−1	Poor
0	None: The appropriateness of the measure was not assessed
1	Minimal/Emerging
2	Adequate
3	Good
4	Excellent
Useful Category	
Fits Organizational Activities	
−1	Poor
0	None: The organizational fit of the measure was not assessed
1	Minimal/Emerging
2	Adequate
3	Good
4	Excellent
Informs Clinical or Organizational Decision-Making	
−1	Poor
0	None: The ability of the measure to inform decision-making was not assessed
1	Minimal/Emerging
2	Adequate
3	Good
4	Excellent
Objective Rating Criteria	
Acceptable Category	
Cost	

(Continued)

Table 3 | Continued

−1	Poor: The measure is extremely costly > $100 per use
0	None: The cost of the measure is unknown
1	Minimal/Emerging: The measure is very costly > $50 but < $100 per use
2	Adequate: The measure is somewhat costly > $1 but < $50 per use
3	Good: The measure is not costly < $1 per use
4	Excellent: The measure is free and in the public domain
Easy Category	
Uses Accessible Language	
−1	Poor: The measure uses language that was only readable by experts in its content
0	None: The measure was not available in the public domain and therefore readability cannot be assessed
1	Minimal/Emerging: The readability of the measure was at a graduate study level
2	Adequate: The readability of the measure was at a college level
3	Good: The readability of the measure was between an 8th and 12th grade level
4	Excellent: The readability of the measure was below an 8th grade level
Assessor Burden (Training)	
−1	Poor: The measure requires an external, expert administrator, with no option to self-train or for a train-the-administrator component
0	None: The training and administration information for the measure is unavailable
1	Minimal/Emerging: The measure requires a train-the-administrator component that is specialized or includes a significant cost
2	Adequate: The measure requires some training, in addition to a manual, and/or supervision/consultation with experts is needed to administer the measure which includes minimal cost (i.e., small consultant fee)
3	Good: The measure includes a manual in order to self-train for administration and the cost for the manual is free or minimal
4	Excellent: The measure requires no training and/or has free automated administration
Assessor Burden (Interpretation)	
−1	Poor: The measure requires an expert to score and interpret, though no entity to whom to send the measure is identified, and no information on handling missing data is provided
0	None: The ease of interpreting the measure cannot be assessed because the measure is not in the public domain
1	Minimal/Emerging: The measure does not include suggestions for interpreting score ranges, no clear cut-off scores, and no instructions for handling missing data
2	Adequate: The measure includes a range of scores with few suggestions for interpreting them but no clear cut-off scores and no instructions for handling missing data
3	Good: The measure includes a range of scores with value labels and cut-off scores, but scoring requires manual calculation and/or additional inspection of response patterns or subscales, and no instructions for handling missing data are provided
4	Excellent: The measure includes clear cut-off scores with value labels, instructions for handling missing data are provided, and calculation of scores is automated or scores can be sent off to an identified entity for calculation with results returned
Length	
−1	Poor: The measure has greater than 200 items
0	None: The measure is not available for use in the public domain
1	Minimal/Emerging: The measure has greater than 100 items but fewer than 200 items
2	Adequate: The measure has greater than 50 items but fewer than 100
3	Good: The measure has greater than 10 items but fewer than 50
4	Excellent: The measure has fewer than 10 items

person would spend looking for basic information online [14,15]; we tested different lengths of time and observed diminishing returns on hits; and (c) only the first page of results was scanned for promising information regarding pragmatic qualities to reflect the average person's willingness to search online search engines for information [16,17]. If a website relating to the measure was located on the first page of results, it was mined for the presence of any of the pragmatic data not already captured in our earlier processes. Then, two research specialists (E.A.N. and K.D.M.) applied the pragmatic rating criteria scale to the data for each of the 60 randomly selected measures. These procedures are similar to those used in our previous work focused on psychometric qualities of measures [6].

RESULTS

Eleven pragmatic properties across the four categories were retained after the refinement process for the final rating criteria: Useful [*n* = 2 items], Compatible [*n* = 1 items], Acceptable [*n* = 3 items], and Easy [*n* = 5 items]. Although not anticipated at first, two distinct types of rating criteria emerged from the synthesis and refinement work: (a) "stakeholder-facing criteria" that require a stakeholder perspective for valid rating, and (b) "objective rating criteria" that could be rated by a team of trained experts. Table 3 shows the full list of stakeholder-facing and objective rating criteria which, when combined with the psychometric rating criteria that were also developed and published under the same grant funding, form the PAPERS [6]. Because these scales were unanticipated, we were not yet able to pilot them with stakeholders.

For the objective rating criteria, we developed nuanced descriptive anchors that were applied in the pilot and refined as needed; the final set of criteria performed well. For instance, when a measure was assessed for its pragmatic quality of "length" (from the Easy domain), a score of −1 (poor) is equal to a measure having greater than 200 items, whereas a score of 4 (excellent) is equal to the measure having fewer than or equal to 10 items.

The range of aggregated scores across the 60 evaluated measures was zero (11 measures) to 15 (only two measures) on a 20-point scale. The mean score was 5.36 and the median was 5. Figure 1 displays the process of locating pragmatic data to inform the objective criteria ratings given to the evaluated implementation measures. No pragmatic data were found for the measure packets or via additional online Google searches for 12 of the 60 measures. Pragmatic data were found in measure packets for 42 measures, and information was found from additional online searches for 5 measures. One measure had pragmatic information from only the additional online search (with no pragmatic information found in the measure packet).

DISCUSSION

This set of studies generated a stakeholder-informed rating system that can be used to evaluate how pragmatic an implementation measure is from a stakeholder's perspective. Following a multistep, iterative process through which we completed a systematic review and engaged stakeholders in selecting measure criteria meaningful to them, we identified four categories and eleven properties of pragmatic measures. These criteria were divided

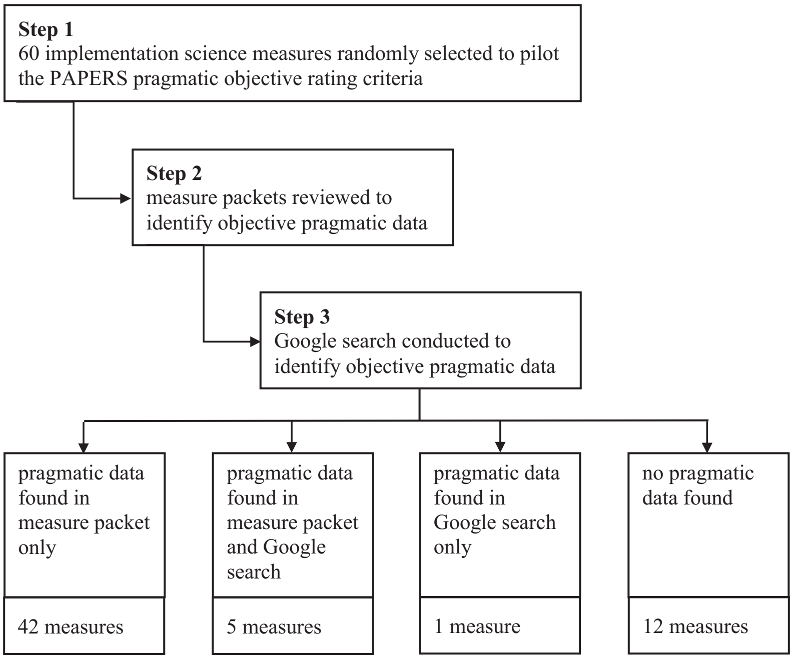
Step 1
60 implementation science measures randomly selected to pilot the PAPERS pragmatic objective rating criteria
Step 2
measure packets reviewed to identify objective pragmatic data
Step 3
Google search conducted to identify objective pragmatic data
pragmatic data found in measure packet only
42 measures
pragmatic data found in measure packet and Google search
5 measures
pragmatic data found in Google search only
1 measure
no pragmatic data found
12 measures

Figure 1 | Process of identifying data to inform pragmatic objective criteria ratings.

into stakeholder-facing (to be administered to stakeholders) and objective domains (can be administered to experts). Although beyond the scope of the current paper and funding, future research efforts should consider including the stakeholder-facing criteria via a survey (web-based) using a Likert scale where 1 represents "strongly disagree" and 5 represents "strongly agree" and assess its interrater reliability. The objective rating criteria was applied to 60 implementation science measures and functioned sufficiently in our pilot study. Combined with our psychometric rating criteria, our PAPERS is ready for further testing and application and has the potential to inform measure development, evaluation, and selection.

Not only did a user-centered design allow us to identify criteria not previously found by other methods (e.g., expert opinion) that were important to stakeholders (e.g., "uses accessible language"), it highlighted that some pragmatic criteria identified by researchers were not, in fact, as important to stakeholders (e.g., "related to a theory or model") [1]. Combining stakeholder-driven pragmatic criteria with more scientific concepts of measure strength (psychometric properties) ensures that we are developing and improving upon measures for their intended audience, and not furthering the research-practice gap within the field that is meant to close it.

An aspect of the pragmatic criteria that unexpectedly emerged was the distinction between properties considered objective (e.g., cost), versus those we termed "stakeholder-facing." As an example of the latter, the term "acceptable" was a priority pragmatic property retained across all study phases and prioritized by our stakeholders. Our ability to clearly define the acceptable domain (i.e., "acceptable," "offers relative advantage over existing methods") across a 6-point rating system was limited given our team's perspective as a team of implementation researchers, rather than practitioner stakeholders whose perspective are grounded in their practice context. Indeed, the individuals best suited to assign a rating for the level of "acceptability" or the "relative advantage" of a particular implementation measure were the stakeholders themselves. As another example, our team did not believe we could judge whether a measure was a fit with organizational activities, believing instead that this could only be determined by stakeholders reflecting on their context. Thus, six of the 11 final pragmatic domains were identified as "stakeholder-facing," and future research is needed to assess the reliability and validity of stakeholder ratings utilizing these criteria. Importantly, given the context-dependent features of at least some of the stakeholder-facing criteria, certain psychometric aspects of pragmatic criteria, such as interrater reliability of the pragmatic ratings among multiple raters, may be especially difficult to apply and evaluate. It will be important for future research endeavors to be explicit and deliberate about which psychometric properties of the pragmatic rating criteria should be assessed.

For the remaining objective pragmatic criteria, we piloted the feasibility of their application using both peer-reviewed and grey literature as data sources. Given the separation of stakeholder-facing vs. objective-facing criteria, it also became clear that the objective criteria (e.g., anchors for the Cost criterion are clearly stated by dollar amounts, etc.) reflected so little subjectivity that interrater reliability was assumed to be quite high, although future testing is still warranted. In addition, despite our assumption that the peer reviewed literature would offer little by way of pragmatic properties, we learned that the majority of the objective criteria are indeed addressed in measure development studies. Counter to our expectations, little benefit was added when drawing from the grey literature. Although this was just a pilot, our evaluation of 60 measures across six domains of implementation constructs and outcomes provide confidence that the objective criteria could be evaluated using peer reviewed literature as the sole data source (i.e., in only one case did we obtain online information when no information was in the empirical literature).

In summary, objective pragmatic rating criteria were combined with psychometric rating criteria to create the PAPERS. PAPERS can be used to assess the quality of evidence for implementation measures reported in systematic reviews, and it can be included in guidelines for reporting on the development and use of implementation measures. When applied in these ways, PAPERS can help researchers and stakeholders select psychometrically and pragmatically strong implementation measures among the many implementation measures currently in existence. Furthermore, this work will inform a measurement development research agenda for the field given the few measures that have been identified as strong in both psychometric and pragmatic properties.

Limitations

There are several noteworthy limitations to this study. First, all stakeholders included in the studies that led to the development of the pragmatic rating criteria represented mental health contexts (including intermediaries whose work was solely focused on mental health agencies, policies, etc.). Although unlikely, it is possible that these stakeholders have very different perspectives about pragmatic implementation measures than do stakeholders from other fields or health more broadly. It will be important for future research to include a purposefully sampled stakeholder group to ensure diversity in context when assessing the reliability and validity of the stakeholder-facing rating criteria.

Second, the term "pragmatic" was used to define the construct as it applied to measurement based on the extant literature of Glasgow and colleagues at the time this set of studies were conducted. It is possible that one of the synonyms identified could be better suited to define the construct; however, given that an existing literature base existed, it seemed the most appropriate label for the construct at the time.

A third limitation is that we did not formally assess certain characteristics of the pragmatic rating criteria, such as known-groups validity. Ultimately, what emerged was that the objective criteria appear to have substantial face validity and to be able to assess known groups would primarily mean piloting the criteria; thus, we chose to pilot the criteria with a larger number of measures instead. The rating criteria may be further strengthened in future research focused on establishing these features, or to formally assess to what degree other forms of psychometric properties may be relevant and acceptable (e.g., interrater reliability), establishing cut-off scores for use in various contexts.

A fourth limitation is that our rating criteria are primarily designed to rate self-report implementation measures; however, our previous work has identified 450+ measures and the vast majority of them are self- or proxy-report, or expert rating scales [4,6]. It remains an empirical question how our PAPERS criteria may respond to different measure formats, such as computer-adapted testing, and future research should consider this.

Finally, we only assessed pragmatic data within each measure's development article during pilot testing; it is possible that other empirical articles including the specified measure informed on its pragmatic properties.

CONCLUSIONS

In sum, this set of studies produced the first user-centered pragmatic measure rating criteria for implementation measures, culminating in the development of the PAPERS. The pragmatic rating criteria can be combined with psychometric rating criteria for improving systematic reviews and informing new measure development. The hope is that practitioners will assume greater independence in integrating measures into their implementation efforts to inform strategy selection, progress monitoring, and outcome evaluations.

Contributions to the literature

- Involving stakeholders at every step of measurement work is critical for bridging the emerging research-practice gap in implementation science.
- We have established the first user-centered pragmatic measures rating criteria.
- The pragmatic measures rating criteria will allow for measure development work to be informed and influenced by stakeholder priorities.
- Establishing the pragmatic rating criteria sets forth a research agenda for evaluating existing measures according to the criteria.

Funding: This study was funded by National Institute of Mental Health (R01 MH106510).

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest: Authors Cameo F. Stanick, Heather M. Halko, Elspeth A. Nolen, Byron J. Powell, Caitlin N. Dorsey, Kayne D. Mettert, Bryan J. Weiner, Melanie Barwick, Luke Wolfenden, Laura J. Damschroder, and Cara C. Lewis have no conflicts of interest to report.

Authors' Contribution: Authors CS, HH, EN, CD, and KM collected data, analyzed data, and contributed to manuscript prep, writing, and editing; BP, BW, and CC analyzed data and contributed to manuscript writing and editing; MB, LW, and LD contributed to manuscript writing and editing and provided consultation on manuscript conceptualization.

Ethical Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent: Informed consent was obtained from all individual participants included in the study.

References

1. Glasgow RE, Riley WT. Pragmatic measures: what they are and why we need them. *Am J Prev Med.* 2013;45(2):237–243.
2. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16(9):606–613.
3. Lewis CC, Fischer S, Weiner BJ, Stanick C, Kim M, Martinez RG. Outcomes for implementation science: an enhanced systematic review of instruments using evidence-based rating criteria. *Implement Sci.* 2015;10:155.
4. Lewis CC, Stanick CF, Martinez RG, et al. The Society for Implementation Research Collaboration Instrument Review Project: a methodology to promote rigorous evaluation. *Implement Sci.* 2015;10:2.
5. Stanick CF, Halko HM, Dorsey CN, et al. Operationalizing the 'pragmatic' measures construct using a stakeholder feedback and a multi-method approach. *BMC Health Serv Res.* 2018;18(1):882.
6. Lewis CC, Mettert KD, Dorsey CN, et al. An updated protocol for a systematic review of implementation-related measures. *Syst Rev.* 2018;7(1):66.
7. Powell BJ, Stanick CF, Halko HM, et al. Toward criteria for pragmatic measurement in implementation research and practice: a stakeholder-driven approach using concept mapping. *Implement Sci.* 2017;12(1):118.
8. Murphy MK, Black NA, Lamping DL, et al. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess.* 1998;2(3):i–iv, 1.
9. Michie S, Abraham C, Eccles MP, Francis JJ, Hardeman W, Johnston M. Strengthening evaluation and implementation by specifying components of behaviour change interventions: a study protocol. *Implement Sci.* 2011;6:10.
10. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs.* 2000;32(4):1008–1015.
11. Powell C. The Delphi technique: myths and realities. *J Adv Nurs.* 2003;41(4):376–382.
12. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci.* 2009;4:50.
13. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm Policy Ment Health.* 2011;38(2):65–76.
14. Brafton BJ. 2017 Content Marketing Benchmark Report. 2017. Available at https://www.brafton.com/blog/strategy/brafton-2017-content-marketing-benchmark-report/
15. Singer G, Norbisrath U, Lewandowski D. Ordinary search engine users carrying out complex search tasks. *J Inf Sci.* 2013;39(3):346–358.
16. Singer G. *Web Search Engines and Complex Information Needs.* Tartu, Estonia: University of Tartu; 2012.
17. Maynes R, Everdell I. *The Evolution of Google Search Results Pages and their Effects on User Behaviour.* USA: Mediative; 2014.