Implementation Research & Practice

Short Report

The psychometric and pragmatic evidence rating scale (PAPERS) for measure development and evaluation

Implementation Research and Practice
1–6

DOI: 10.1177/26334895211037391
journals.sagepub.com/home/irp
SAGE

Cara C Lewis[1,2], Kayne D Mettert[1], Cameo F Stanick[3], Heather M Halko[4], Elspeth A Nolen[5], Byron J Powell[6] and Bryan J Weiner[5,7]

Abstract

To rigorously measure the implementation of evidence-based interventions, implementation science requires measures that have evidence of reliability and validity across different contexts and populations. Measures that can detect change over time and impact on outcomes of interest are most useful to implementers. Moreover, measures that fit the practical needs of implementers could be used to guide implementation outside of the research context. To address this need, our team developed a rating scale for implementation science measures that considers their psychometric and pragmatic properties and the evidence available. The Psychometric and Pragmatic Evidence Rating Scale (PAPERS) can be used in systematic reviews of measures, in measure development, and to select measures. PAPERS may move the field toward measures that inform robust research evaluations and practical implementation efforts.

Keywords

Measures, psychometric, pragmatic, rating scale, implementation

Implementation science is uniquely poised to supply measures to guide implementation of evidence-based interventions. That is, practitioners could use measures to plan implementation efforts, monitor implementation processes, and evaluate implementation outcomes. For example, mounting evidence suggests that implementation leadership can drive effective implementation (Aarons et al., 2014; Aarons et al., 2016; Farahnak et al., 2020). By administering measures like the survey by Aarons et al. (2014) on implementation leadership, organizations could more efficiently identify and develop leaders who are proactive, knowledgeable, supportive, and perseverant, which are empirically identified qualities critical to implementation success. To carry on this example, in the absence of measures to offer this precise focus, leaders may be identified who are not equipped to guide implementation of an evidence-based intervention through to sustainment.

To have confidence in the results of our measures, psychometric evidence in a variety of contexts with diverse populations is key. That is, measures themselves are not "validated" but evidence of their validity is established through empirical uses, ideally through psychometric evaluations. Psychometric properties of particular practical relevance include sensitivity to change (i.e., the degree to

[1]Kaiser Permanente Washington Health Research Institute, USA
[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Harborview Medical Center, USA
[3]Hathaway-Sycamores Child and Family Services, USA
[4]University of Montana, USA
[5]Department of Global Health, University of Washington, USA
[6]Brown School and School of Medicine, Washington University in St. Louis, USA
[7]Department of Health Services, University of Washington, Seattle, WA, USA

Corresponding author:
Cara Lewis, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, USA.
Email: cara.c.lewis@kp.org

Table 1. Psychometric and pragmatic evidence rating scale (PAPERS).

Reliability—Internal Consistency
−1 Poor (P): Cronbach's α values of < 0.50
0 None (N): Internal consistency measures are not applicable for this instrument OR Classical Test Theory anchors are not appropriate (results reported using Item Response Theory) OR α values are not yet available for the full measure scale or any associated subscales.
1 Minimal/Emerging (M): Cronbach's α values = 0.50–0.69
2 Adequate (A): Cronbach's α values of = 0.70–0.79
3 Good (G): Cronbach's α values of = 0.80–0.89
4 Excellent (E): Cronbach's α values of ≥ 0.90
Note: When only subscale α values are given, provide all and apply the 'worst score counts' rule.
Construct Validity—Convergent
−1 Poor: Cohen's $d \leq 0.10$
0 None (N): Convergent validity measures are not applicable for this instrument OR convergent validity was not assessed.
1 Minimal/Emerging: 0.10 < Cohen's $d \leq 0.20$
2 Adequate: 0.20 < Cohen's $d \leq 0.50$
3 Good: 0.50 < Cohen's $d \leq 0.80$
4 Excellent: Cohen's $d > 0.80$
Note: If Pearson's r is given, use the effect size calculator to calculate Cohen's *d*. https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html
Also, note that these criteria also apply to comparisons between subscales.
Construct Validity—Discriminant
−1 Poor: Cohen's $d > 0.80$
0 None (N): Discriminant validity measures are not applicable for this instrument OR discriminant validity was not assessed.
1 Minimal/Emerging: 0.50 < Cohen's $d \leq 0.80$
2 Adequate: 0.20 < Cohen's $d \leq 0.50$
3 Good: 0.10 < Cohen's $d \leq 0.20$
4 Excellent: Cohen's $d \leq 0.10$
Note: If Pearson's r is given, use the effect size calculator to calculate Cohen's *d*. https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html
Also, note that these criteria also apply to comparisons between subscales.
Construct Validity—Known Groups
Categories: Demographics, Roles/Professions, Programs/Treatments, Organizations, Intervention Conditions
−1 Poor (P): Known-groups validity failed to be detected.
0 None (N): Known-groups validity not yet tested.
1 Minimal/Emerging (M): Statistically significant difference between groups detected, but no hypothesis tested
2 Adequate (A): Two or more statistically significant difference between groups detected, but no hypotheses tested
3 Good (G): Statistically significant difference between groups detected AND hypothesis tested
4 Excellent (E): Two or more statistically significant differences between groups detected AND hypotheses tested
Criterion Validity—Predictive
Evidence of correlation (Pearson's r) between instrument and scores on another test (measuring a distinct construct of interest or outcome) administered at some point in the future.
−1 Poor (P): Pearson's r < 0.10
0 None (N): Predictive validity not tested.
1 Minimal/Emerging (M): Pearson's r = 0.10–0.29
2 Adequate (A): Pearson's r = 0.30–0.49
3 Good (G): Pearson's r = 0.50–0.69
4 Excellent (E): Pearson's r > 0.70
Note: If unstandardized regression coefficients (betas) are reported, use the effect size calculator to translate them into Pearson's r values and follow the same rules as above.
https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-R7.php
&
If discriminant function analysis is reported, use the measure of variance explained. Anchors for this can be found in the "Structural Validity" section.

(Continued)

Table 1. (Continued)

Criterion Validity—Concurrent
Evidence of correlation (Pearson's r) between instrument and scores on another test (measuring a distinct construct of interest or outcome) administered at the same point in time.
−1 Poor (P): Pearson's r < 0.10
0 None (N): Concurrent validity not tested.
1 Minimal/Emerging (M): Pearson's r = 0.10–0.29
2 Adequate (A): Pearson's r = 0.30–0.49
3 Good (G): Pearson's r = 0.50–0.69
4 Excellent (E): Pearson's r > 0.70
Note: If unstandardized regression coefficients (betas) are reported, use the effect size calculator to translate them into Pearson's r values and follow the same rules as above.
https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-R7.php
&
If discriminant function analysis is reported, use the measure of variance explained. Anchors for this can be found in "Structural Validity" section.
Dimensionality—Structural Validity
Normed Fit Index = NFI; Incremental Fit Index = IFI Goodness of Fit Index = GFI; Tucker–Lewis Index = TLI Comparative Fit Index = CFI; Relative Noncentrality Fit Index = RNI Standardized RMR = SRMR; Root Mean Square Error of Approximation = RMSEA Weighted Root Mean Residual = WRMR
−1 Poor (P): The sample consisted of less than 5 times the number of items AND exploratory factor analysis explained < 25% of variance OR NFI OR IFI OR GFI OR TLI OR CFI OR RNI $\leq$ 0.88 OR SRMR OR RMSEA = X $\geq$ 0.10 OR WRMR $\geq$ 0.92
0 None (N): No exploratory or confirmatory factor analysis has yet been performed, nor have any Item Response Theory (IRT) tests of (uni-) dimensionality have been conducted OR analysis has been conducted but percent variance is unexplained and cannot be calculated OR only principal components analysis has been conducted.
1 Minimal/Emerging (M): The sample consisted of 5 times the number of items AND exploratory factor analysis explained < 25% of variance OR NFI OR IFI OR GFI OR TLI OR CFI OR RNI = 0.88 <X $\leq$ 0.90 OR SRMR OR RMSEA = 0.08 $\leq$X < 0.10 OR WRMR = 0.90 $\leq$X < 0.92
2 Adequate (A): The sample consisted of 5 times the number of items but is less than 100 in total AND an exploratory factor analysis explained < 50% of variance OR NFI OR IFI OR GFI OR TLI OR CFI OR RNI = 0.90 <X $\leq$ 0.95 OR SRMR OR RMSEA = 0.05 $\leq$X < 0.08 OR WRMR = 0.85 $\leq$X < 0.90
3 Good (G): The sample consisted of 5 times the number of items and is greater than or equal to 100 in total OR the sample consisted of 5–7 times the number of items but is less than 100 in total AND in either case exploratory factor analysis explained < 50% of variance OR NFI OR IFI OR GFI OR TLI OR CFI OR RNI = 0.95 <X $\leq$ 0.97 OR SRMR OR RMSEA = 0.03 $\leq$X < 0.05 OR WRMR = 0.83 $\leq$X < 0.85
4 Excellent (E): The sample consisted of 7 times the number of items and is greater than 100 in total AND an exploratory factor analysis explained > 50% of variance OR NFI OR IFI OR GFI OR TLI OR CFI OR RNI > 0.97 OR SRMR OR RMSEA = < 0.03 OR WRMR < 0.83
Note: If multiple indices are given and they fall within differing rating anchors, use the mode score (three "good" ratings, one excellent rating, one poor rating → rated as "good.")
Responsiveness
Standardized Response Mean = SRM
−1 Poor (P): SRM < 0.10 OR Pearson's r < 0.10
0 None (N): The instrument has not been administered both pre- and post-implementation to evaluate sensitivity to change.

(Continued)

Table 1. (Continued)

1	Minimal/Emerging (M): SRM = 0.10–0.19 OR Pearson's r = 0.10–0.29
2	Adequate (A): SRM = 0.20–0.49 OR Pearson's r = 0.30–0.49
3	Good (G): SRM = 0.50–0.79 OR Pearson's r = 0.50–0.69
4	Excellent (E): SRM > 0.80 OR Pearson's r > 0.70
Norms	
−1	Poor (P): Measures of central tendency and distribution for the total score (and subscales if relevant) based only on a very small ($n < 50$) sample are available.
0	None (N): Norms not yet available.
1	Minimal/Emerging (M): Measures of central tendency and distribution for the total score (and subscales if relevant) based only on a small ($n = 50$–99) sample are available.
2	Adequate (A): Measures of central tendency and distribution for the total score (and subscales if relevant) based only on a small ($n = 100$–299) sample are available.
3	Good (G): Measures of central tendency and distribution for the total score (and subscales if relevant) based on a medium ($n = 300$–499) sample are available.
4	Excellent (E): Measures of central tendency and distribution for the total score (and subscales if relevant) based on a large ($n \geq 500$) sample are available.
Cost	
−1	Poor (P): The measure is extremely costly, ≥$100 per use.
0	None (N): The cost of the measure is unknown.
1	Minimal/Emerging (M): The measure is very costly ≥$50 but < $100 per use.
2	Adequate (A): The measure is somewhat costly ≥$1 but < $50 per use.
3	Good (G): The measure is not costly at < $1 per use.
4	Excellent (E): The measure is free and in the public domain.
Language	
−1	Poor (P): The measure used language that was only readable by experts in its content.
0	None (N): The measure was not available in the public domain and therefore the readability cannot be assessed.
1	Minimal/Emerging (M): The readability of the measure was at a graduate study level (range: 17.0 and above).
2	Adequate (A): The readability of the measure was at a college level (range: 13.0–16.99).
3	Good (G): The readability of the measure was between an 8th and 12th grade level (range: 8.0–12.99).
4	Excellent (E): The readability of the measure was at or below an 8th grade level (range: 7.9 and below).
Microsoft Word Readability Test Instructions: File > Options > Proofing > Check the box "show readability statistics." Readability results will show up at the end of a Spell Check (Review tab).	
Assessor Burden (Ease of training)	
−1	Poor (P): The measure requires an external, expert administrator, with no option to self-train or for a train-the-administrator component.
0	None (N): The training and administration information for the measure is unavailable.
1	Minimal/Emerging (M): The measure requires a train-the-trainer to administer component that is specialized or includes a significant cost (≥$100).
2	Adequate (A): 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) (≥$50 but <$100).
3	Good (G): The measure includes a manual in order to self-train for administration and the cost for the manual is free or minimal (<$50 but not free).
4	Excellent (E): The measure requires no training and/or has free automated administration.
Assessor Burden (Easy to Interpret)	
−1	Poor (P): 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 (N): The ease of interpreting the measure cannot be assessed because the measure is not in the public domain.
1	Minimal/Emerging (M): The measure does not include suggestions for interpreting score ranges, no clear cut-off scores, and no instructions for handling missing data.
2	Adequate (A): 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.

(Continued)

Table 1. (Continued)

3	Good (G): 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 (E): 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 (P): The measure has >200 items.
0	None (N): The measure is not available for use in the public domain.
1	Minimal/Emerging (M): The measure has >100 items but ≤200 items.
2	Adequate (A): The measure has >50 items but ≤ 100 items.
3	Good (G): The measure has >10 items but ≤50 items.
4	Excellent (E): The measure has ≤10 items.

which a measure is designed to detect change over time) and predictive validity (i.e., the degree to which a measure is designed to detect changes in theorized outcomes of interest) as implementers can then monitor the progress and impact of their work.

However, to be taken up for practical use, measures need to also have pragmatic properties. Glasgow and Riley (Glasgow & Riley, 2013) contend that pragmatic measures are critical for implementation, to address stakeholder issues, and drive quality improvement. Pragmatic properties extend beyond psychometric properties to include features like actionability and low burden. In recent years, implementation scientists developing measures seem to be striving toward the development of pragmatic measures, and there are now several brief, actionable tools for the field to deploy (e.g., (Weiner et al., 2017).

Although several rating scales exist to assess the psychometric strength of measures, such as Hunsley and Mash's criteria for evidence-based assessment (Hunsley & Mash, 2008) and the COnsensus-based Standards for the selection of health status Measurement Instruments (COSMIN) checklist for evaluating the methodological quality of studies on measurement properties (Mokkink et al., 2010), none existed to support evaluation of nuanced properties relevant to the complex evaluations occurring in implementation science. Our team was funded by the National Institute of Mental Health (R01MH106510) to develop and refine a psychometric evidence rating scale (Lewis et al., 2018) for application in our systematic reviews of 47 implementation science constructs featured in this special collection. Our team also aimed to co-develop a pragmatic properties rating scale with implementation stakeholders (Powell et al., 2017; Stanick et al., 2019) to apply in tandem with our psychometric evidence rating scale in a subset of our systematic reviews (Powell et al., 2017), and to inspire measure development with pragmatic properties at the fore. We present here the Psychometric and Pragmatic Evidence Rating Scale (PAPERS; Table 1). All or portions of PAPERS were applied across the seven systematic reviews in this special collection entitled, "Systematic Reviews of Methods to Measure Implementation Constructs." Its uses extend beyond the collection (Allen et al., 2020; Gabriella et al., 2021; Khadjesari et al., 2020), however, to application in future systematic reviews and inform the development and selection of implementation measurement.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The psychometric and pragmatic evidence rating scale (PAPERS) for measure development and evaluations. Funding for this study came from the National Institute of Mental Health, awarded to Dr. Cara C. Lewis as principal investigator (R01MH106510). Dr. Lewis is both an author of this manuscript and editor of the journal, *Implementation Research and Practice*. Due to this conflict, Dr. Lewis was not involved in the editorial or review process for this manuscript.

Funding

This work was supported by the National Institute of Mental Health (NIMH) "Advancing implementation science through measure development and evaluation" [1R01MH106510], awarded to Dr. Cara C. Lewis as principal investigator.

ORCID iDs

Cara C Lewis https://orcid.org/0000-0001-8920-8075
Kayne D Mettert https://orcid.org/0000-0003-1750-7863
Cameo F Stanick https://orcid.org/0000-0002-6076-3726
Byron J Powell https://orcid.org/0000-0001-5245-1186

References

Aarons, G. A., Ehrhart, M. G., & Farahnak, L. R. (2014). The implementation leadership scale (ILS): Development of a brief measure of unit level implementation leadership. *Implementation Science*, *9*(1), 45. https://doi.org/10.1186/1748-5908-9-45

Aarons, G. A., Green, A. E., Trott, E., Willging, C. E., Torres, E. M., Ehrhart, M. G., & Roesch, S. C. (2016). The roles of

system and organizational leadership in system-wide evidence-based intervention sustainment: A mixed-method study. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(6), 991–1008. https://doi.org/10.1007/s10488-016-0751-4

Allen, P., Pilar, M., Walsh-Bailey, C., Hooley, C., Mazzucca, S., Lewis, C. C., Mettert, K. D., Dorsey, C. N., Purtle, J., Kepper, M. M., Baumann, A. A., & Brownson, R. C. (2020). Quantitative measures of health policy implementation determinants and outcomes: A systematic review. *Implementation Science*, *15*(1), 47. https://doi.org/10.1186/s13012-020-01007-w

Farahnak, L. R., Ehrhart, M. G., Torres, E. M., & Aarons, G. A. (2020). The influence of transformational leadership and leader attitudes on subordinate attitudes and implementation success. *Journal of Leadership & Organizational Studies*, *27*(1), 98–111. https://doi.org/10.1177/1548051818824529

McLoughlin, G. M., Allen, P., Walsh-Bailey, C., & Brownson, R. C. (2021). A systematic review of school health policy measurement tools: Implementation determinants and outcomes. *Implementation Science Communications*, *2*(1), 67. https://doi.org/10.1186/s43058-021-00169-y. PMID: 34174969; PMCID: PMC8235584.

Glasgow, R. E., & Riley, W. T. (2013). Pragmatic measures: What they are and why we need them. *American Journal of Preventive Medicine*, *45*(2), 237–243. https://doi.org/10.1016/j.amepre.2013.03.010

Hunsley, J., & Mash, E. J. (2008). *A guide to assessments that work*. [https://doi.org/10.1093/med:psych/9780195310641.001.0001]. Oxford University Press. https://doi.org/10.1093/med:psych/9780195310641.001.0001

Khadjesari, Z., Boufkhed, S., Vitoratou, S., Schatte, L., Ziemann, A., Daskalopoulou, C., Uglik-Marucha, E., Sevdalis, N., & Hull, L. (2020). Implementation outcome instruments for use in physical healthcare settings: A systematic review. *Implementation Science*, *15*(1), 66. https://doi.org/10.1186/s13012-020-01027-6

Lewis, C. C., Mettert, K. D., Dorsey, C. N., Martinez, R. G., Weiner, B. J., Nolen, E., Stanick, C., Halko, H., & Powell, B. J. (2018). An updated protocol for a systematic review of implementation-related measures. *Systematic Reviews*, *7*(1), 66. https://doi.org/10.1186/s13643-018-0728-3

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Quality of Life Research*, *19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Powell, B. J., Stanick, C. F., Halko, H. M., Dorsey, C. N., Weiner, B. J., Barwick, M. A., Damschroder, L. J., Wensing, M., Wolfenden, L., & Lewis, C. C. (2017). Toward criteria for pragmatic measurement in implementation research and practice: A stakeholder-driven approach using concept mapping. *Implementation Science*, *12*(1), 118. https://doi.org/10.1186/s13012-017-0649-x

Stanick, C. F., Halko, H. M., Nolen, E. A., Powell, B. J., Dorsey, C. N., Mettert, K. D., Weiner, B. J., Barwick, M., Wolfenden, L., Damschroder, L. J., & Lewis, C. C. (2019). Pragmatic measures for implementation research: Development of the psychometric and pragmatic evidence rating scale (PAPERS). *Translational Behavioral Medicine*, *11*(1), 11–20. https://doi.org/10.1093/tbm/ibz164

Weiner, B. J., Lewis, C. C., Stanick, C., Powell, B. J., Dorsey, C. N., Clary, A. S., Boynton, M. H., & Halko, H. (2017). Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science*, *12*(1), 108. https://doi.org/10.1186/s13012-017-0635-3