

Development and Validation of a Measure of Primary Care Behavioral Health Integration

Rodger S. Kessler, PhD, ABPP
University of Vermont College of Medicine

Andrea Auxier, PhD
New Directions Behavioral Health, Kansas City,
Missouri

Juvena R. Hitt, BS
University of Vermont College of Medicine

C. R. Macchi, PhD
Arizona State University

Daniel Mullin, PsyD, MPH
University of Massachusetts Medical School

Constance van Eeghen, DrPH and
Benjamin Littenberg, MD
University of Vermont College of Medicine

Introduction: We developed the Practice Integration Profile (PIP) to measure the degree of behavioral health integration in clinical practices with a focus on primary care (PC). Its 30 items, completed by providers, managers, and staff, provide an overall score and 6 domain scores derived from the Lexicon of Collaborative Care. We describe its history and psychometric properties. **Method:** The PIP was tested in a convenience sample of practices. Linear regression compared scores across integration exemplars, PC with behavioral services, PC without behavioral services, and community mental health centers without PC. An additional sample rated 4 scenarios describing practices with varying degrees of integration. **Results:** One hundred sixty-nine surveys were returned. Mean domain scores ran from 49 to 65. The mean total score was 55 (median 58; range 0–100) with high internal consistency (Cronbach's alpha = .95). The lowest total scores were for PC without behavioral health (27), followed by community mental health centers (44), PC with behavioral health (60), and the exemplars (86; $p < .001$). Eleven respondents rerated their practices 37 to 194 days later. The mean change was + 1.5 (standard deviation = 11.1). Scenario scores were highly correlated with the degree of integration each scenario was designed to represent (Spearman's $\rho = -0.71$; $P = 0.0005$). **Discussion:** These data suggest that the PIP is useful, has face, content, and internal validity, and distinguishes among types of practices with known variations in integration. We discuss how the PIP may support practices and policymakers in their integration efforts and researchers assessing the degree to which integration affects patient health outcomes.

Keywords: integrated care, measurement, methods, primary care

Rodger S. Kessler, PhD, ABPP, Department of Family Medicine, University of Vermont College of Medicine; Andrea Auxier, PhD, New Directions Behavioral Health, Kansas City, Missouri; Juvena R. Hitt, BS, Department of Family Medicine, University of Vermont College of Medicine; C. R. Macchi, PhD, Integrated Behavioral Healthcare, Arizona State University; Daniel Mullin, PsyD, MPH, Department of Family Medicine and Community Health, University of Massachusetts Medical School; Constance van Eeghen, DrPH and Benjamin Littenberg, MD, Department of Family Medicine, University of Vermont College of Medicine.

Rodger S. Kessler, Andrea Auxier, Juvena R. Hitt, C. R. Macchi, Daniel Mullin, Constance van Eeghen, and Benjamin Littenberg participated in the study design and coordination and helped to draft the manuscript. Additionally, B.L. performed the statistical analysis. All authors read and approved the final manuscript.

Correspondence concerning this article should be addressed to Rodger S. Kessler, PhD, ABPP, Department of Family Medicine, University of Vermont College of Medicine, 89 Beaumont Avenue, Given Courtyard South S456, Burlington, VT 05405. E-mail: rodger.kessler@uvm.edu

Behavioral Health (BH, defined here as mental health, substance abuse and health behavior services) is critically important to maintaining and improving health in Primary Care (PC) settings. Although some PC practices have long been able to integrate BH services, broad, effective dissemination continues to be challenging (Dickinson, 2015). Efforts toward the goal of delivering BH services to all who need them have been hampered by confusion over what services should be included and how they should be integrated into PC. Substantial progress on this front was made with the appearance of the *Lexicon for Behavioral Health and Primary Care Integration* (Peek & the National Integration Academy Council, 2013) that proposed a common language for describing multiple domains of integrated health care.

Although several checklists of collaboration and integration are available, there is no validated measure for describing or measuring the degree of BH integration in any particular PC setting. This limits the abilities of researchers, providers, managers, and policymakers to assess the value of Integrated Behavioral Health, make decisions about resource allocation, design and manage efforts to achieve and maintain it, and reward its achievement. This paper describes the development and validation of the Practice Integration Profile (PIP), a self-administered, Web-based survey that allows providers, staff, and managers to assess their own practices' progress toward an idealized goal of "fully integrated behavioral health services." Moreover, the PIP generates data to inform research about the effectiveness of varying degrees of integration.

As defined by the Agency for Health care Research and Quality (AHRQ) Integration Academy, integrated care is

A practice team of primary care and behavioral health clinicians working together with patients and families, using a systematic and cost-effective approach to provide patient-centered care for a defined population. This care may address mental health and substance abuse conditions, health behaviors (including their contribution to chronic medical illnesses), life stressors and crises, stress-related physical symptoms, and ineffective patterns of health care utilization. (Peek & the National Integration Academy Council, 2013)

Integrated care, supported by a growing body of evidence, has become increasingly mainstream (Miller, 2015). Examples of behavioral

health and primary care services in family medicine residency practices have existed for more than 30 years (Blount & Miller, 2009). The Department of Veterans Affairs and Department of Defense, with their unique population and financial structures, have been leaders in integration (Hunter, Goodie, Dobmeyer, & Dorrance, 2014). Innovative practice organizations focused on underserved populations, such as the South Central Foundation in Alaska and Cherokee Health Systems in Tennessee, have developed financially sustainable integrated care models in their communities (Cohen et al., 2015).

Multiple challenges hinder systematic, integrated care. No single set of metrics exists to guide program implementation or to evaluate the Triple Aim outcomes of improved patient experience, better outcomes, and lowered cost of care (Institute for Healthcare Improvement, 2015). If integration is to generate accelerated implementation, and if research focused on integrated Behavioral Health and identification best practices is to occur, a validated measure of what is being done in integrated practices is needed. Therefore, we sought to develop and validate a measure of the degree to which practices achieved an idealized state of integration.

Method

The Instrument

The approaches and parameters delineated in the AHRQ Lexicon (Peek & the National Integration Academy Council, 2013) served as the theoretical foundation of a new measure of integrated care, the Practice Integration Profile (PIP, formerly the Vermont Integration Profile (Kessler et al., 2015). The authors of the PIP began with a detailed review of the Lexicon's defining clauses, alternatives, and parameters and then developed questions organized into six domains of integrated care.

Pilot testing demonstrated that initial versions of some of the questions were ambiguous, and some of the domains overlapped enough to be combined. The current version of the PIP has 30 questions. Most of the questions have the stem "In our practice . . ." followed by a practice characteristic (such as ". . . we use registry tracking for patients with identified BH issues"), an example ("Insom-

nia registry”), a definition (“Numerator = # of patients in BH registries; Denominator = # of patients with BH needs”), and five response options. The options include: *Never* (0%), *Sometimes* (1–33%), *Often* (34–66%), *Frequently* (67–99%), and *Always* (100%). Each of the possible responses is assigned a score near the midpoint of its stated range: 0, 25, 50, 75 and 100%. (Two of the questions have different response options. The full instrument appears in the Appendix).

The PIP is organized into six domains. *Practice Workflow* includes the policies and procedures that ensure the organizational structure to support consistent delivery of evidence- services to patients in need. *Work-space Arrangement and Infrastructure* addresses the physical proximity and use of shared medical records. *Integration Methods (Shared Care)* covers the type and degree of interactions among medical and behavioral providers. *Case Identification* specifies the practice’s procedures for screening and identifying patients who need BH services. *Patient Engagement* captures the ability of the practice to initiate treatment, involve the patient in developing and delivering the care, and provide support to the patient through ongoing management and follow-up. The domains contain between two and nine questions each. They are scored as the average of their item scores. All scores can run from 0 (*least degree of integration*) to 100 (*greatest degree of integration*). The *Total Integration Score* is the unweighted numeric average of the six domains.

The PIP was presented to respondents by email invitation and administered via REDCap (Harris et al., 2009) a secure online survey system that automatically scores the responses and provides tabular and graphical feedback to the respondent comparing their scores to others. When administered in this fashion, there are no missing values. However, if the PIP is administered by a mechanism that allows skipped items or missing values (such as paper and pencil), the scoring algorithm calls for using the average of all the responses available for each domain as long as there are at least two valid responses in that domain. The Total Integration Score is not calculated unless all six domains are available.

Scenario Studies

Prior to field testing, a sample of five raters used the measure to evaluate four practice scenarios describing hypothetical primary care practices with varying degrees of Behavioral Health Integration. The scenarios were each approximately one page long and described the physical arrangement of the practice, staff, the type of services offered and other information needed to assess the degree of BH integration. Raters were experienced primary care or behavioral health providers. Each rater ranked each scenario from *most integrated* (1) to *least integrated* (4) and completed a PIP for each one. We hypothesized that if the PIP reflected their gestalt judgments, the Total Integration Scores (and to a lesser degree, the domain scores) would correlate with their rankings. We tested this with a nonparametric correlation coefficient (Spearman’s rho; Spearman, 1904).

Field Testing

The PIP was then tested in a convenience sample of primary care and BH practices recruited from email broadcasts to relevant listserves, national webinars, and national meetings. It was completed by physicians, BH clinicians, managers, and staff within the practices. Respondents were eligible if their practice provided Primary Care or Community Mental Health services with or without integrated BH and medical services.

In addition to the PIP, each respondent provided the name and location of their practice, their role (PC provider, BHC, manager, staff, or student), practice type, specialty, and number of providers. We asked respondents to base their responses on their personal knowledge of the practice and did not require that they measure any of the items with exactitude. We divided the practices into four levels of integration. Those with no behavioral or mental health clinicians were expected to have the lowest PIP scores, followed by Community Mental Health Centers (CMHC), and then PC practices. Based on reports by the Agency for Health care Research and Quality (Cohen et al., 2015), eight PC practices were identified as “exemplars” representing the most advanced examples of BH integration and were expected to have the highest PIP scores.

A small subset of respondents was asked to repeat the assessment weeks after their initial report.

Respondents received no compensation. The protocol was reviewed by the University of Vermont IRB and assessed as exempt from human subjects research regulations.

Analysis

We used Cronbach's alpha to assess the internal consistency of each domain scale and the Total Integration Score in the sample of 169 responses. (Bland & Altman, 1997; Cronbach, 1951). Analysis of variance (ANOVA) and linear regression were used to compare total and domain scores across four types of practice (PC without behavioral services, community mental health centers without PC, PC with some behavioral services, and exemplars) while controlling for other practice characteristics. We used Spearman's rho, a nonparametric method, to assess correlation (Spearman, 1904), and Cuzick's rank sum test to assess trends in scores across levels of integration (Cuzick, 1985). Graphical tools included Tukey Box Plots (Tukey, 1977) for distributions of continuous variables and paired-point scatter plots for bivariate associations.

Results

One hundred sixty-nine surveys were completed by staff at 152 practices in 35 states. The mean number of responses per practice was 1.1 (range 1 to 3). The respondents include: 61 BHCs, 34 PCPs, 67 managers and seven student BHCs. The practices serve inner city (15), urban (54), suburban (32), rural (46), and frontier (5) communities. Thirty-six are Community Health

Centers and 23 are Community Mental Health Centers (CMHCs). Fifty-nine are Family Medicine, 18 Internal Medicine, two Pediatric, two Obstetric, and 12 multispecialty practices. The practices tend to be large with 135 reporting over 10 providers and only eight having fewer than six providers.

Of 169 collected surveys, 90% were completed in full. User reports suggest that completing the PIP is approximately a 10-min task. There were no significant differences between types of respondents- physicians, behavioral health clinicians, administrators, or other categories of rater.

The mean of the 169 Total Integration Scores was 55 (standard deviation 20) with median 58 and range from 0 to 100. The median domain scores were Workflow (54), Clinical Services (67), Workspace (75), Shared Care & Integration (50), Case Identification (50), and Patient Engagement (50; see Table 1). The distribution of scores used the full range of potential values (0–100) for each domain and tended to be symmetrical (see Figure 1). The exception is the Workspace domain which has only two items.

Internal Consistency

The scale reliability or internal consistency of each domain scale, expressed as Cronbach's alpha, ranged from 0.52 to 0.91. The internal consistency of the Total Integration Score was $\alpha = .95$ (see Table 1).

Discrimination Among Levels of Integration

The average Total Integration Score was 27 for Non-Behavior Health Clinician (Non-BHC) practices, 44 for Community Mental Health

Table 1
Practice Integration Profile Domain Scores

Domain	<i>k</i>	Mean	<i>SD</i>	Minimum	25th percentile	Median	75th percentile	Maximum	α
Workflow	6	53	21	0	38	54	67	100	.82
Services	9	62	24	0	44	67	81	100	.91
Workspace	2	65	33	0	38	75	100	100	.52
Shared care	4	49	27	0	25	50	69	100	.87
Identification	5	51	26	0	25	50	70	100	.88
Engagement	4	49	22	0	31	50	63	100	.80
Total	30	55	21	0	40	58	71	100	.95

Note. *k* = number of items; *SD* = standard deviation; α = Cronbach's α .

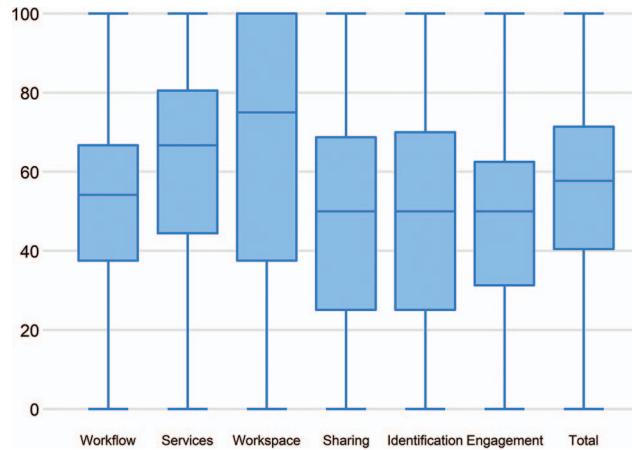


Figure 1. Distribution of Practice Integration profile scores by domain. Each box-and-whisker plot represents the distribution of a domain score for all respondents. Each box runs from the 25th to the 75th percentile of scores with the median drawn as a band across the middle of the box. The whiskers extend to the minimum and maximum scores. There were no outliers. See the online article for the color version of this figure.

Centers (CMHCs), 60 for Primary Care practices, and 86 for Exemplars ($F = 20.2$ by ANOVA; $p < .0001$). Similar differences were observed in the median values of the four types of practices (see Figure 2). For the individual domains, in nearly every case, the scores increased monotonically as predicted from Non-BHC to CMHC to PC to Exemplar (see Table

2). The only exception was in the Case Identification domain, where CMHCs had somewhat lower scores than the No Behaviorist practices. This difference was not significant (43 vs. 37; $p = .45$).

In linear regression, the PIP yielded significantly different Total Integration scores among all four practices types with $p < .001$ for all

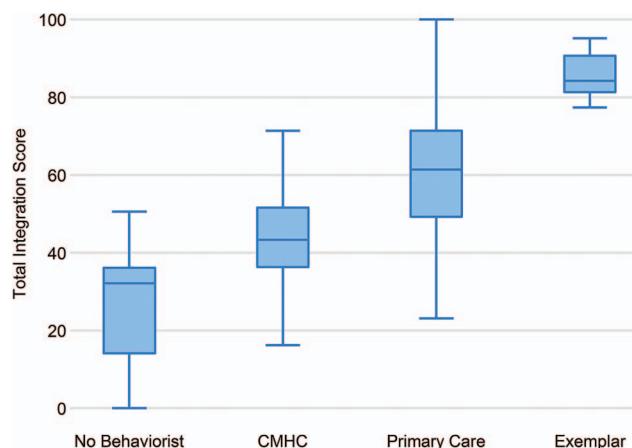


Figure 2. Practice Integration Profile total score by level of integration (discrimination). Each box-and-whisker plot represents the distribution of Total Integration Scores for a subgroup of practices. Each box runs from the 25th to the 75th percentile of scores with the median drawn as a band across the middle of the box. The whiskers extend to the minimum and maximum scores. There were no outliers. See the online article for the color version of this figure.

Table 2
Average Domain Scores by Level of Integration

Level	Respondents	Practices	Average domain scores							
			Workflow	Services	Workspace	Shared care	Identification	Engagement	Total	
Non-BHC	20	20	31	23	21	18	43	27	27	
CMHC	25	22	54	56	38	32	37	49	44	
Primary care	114	102	55	67	77	56	53	51	60	
Exemplar	10	8	80	94	98	86	83	73	86	
All practices	169	152	53	62	65	49	51	49	55	

Note. Non-BHC = Practice with no Behavioral Health Clinicians. CMHC = Community Mental Health Center. Within each domain, the trend across levels is statistically significant ($p < .001$) by the Cuzick nonparametric test of trend (Cuzick, 1985).

comparisons, demonstrating ability to discriminate across all levels of integration. Expanding the model to control for potential confounding by practice size, practice location and respondent type had little effect on the coefficients for each level of integration. A similar pattern of minimal change when controlling for potential confounders was observed in all the domains (see Table 3).

Intrarater Consistency Over Time (Test-Retest Reliability)

Among 11 subjects who repeated the survey 37 to 194 days later (median 48), the mean change in Total Integration Score was +1.5 of 100 (95% confidence interval = -5.0, +8.0)

with a range from -19 to + 23, providing evidence of good test-retest reliability. There was no association between the time between assessments and the change in the total integration score. In linear regression, the coefficient on days was 0.07 (95% confidence interval = -0.10, +0.25; $P = 0.38$; see Figure 3). The individual domains had somewhat larger changes. See Table 4.

Within Practice Agreement

Fifteen practices had multiple respondents including two practices with three respondents ($n = 32$ respondents). The Total Integration scores among respondents from the same practices appears in Figure 4. The mean difference in Total

Table 3
Effect of Potential Confounders on the Domain Scores by Level of Integration

Domain	Workflow		Services		Workspace		Shared care		Identification		Engagement		Total		
	Model	A	B	A	B	A	B	A	B	A	B	A	B	A	B
Level of integration															
Non-BHC	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
CMHC	22 [†]	22 [†]	33 [†]	35 [†]	17 [*]	15 [*]	14 [*]	15 [*]	-6	-6	21 [†]	21 [†]	17 [†]	17 [†]	
Primary care	24 [†]	22 [†]	44 [†]	40 [†]	56 [†]	51 [†]	38 [†]	37 [†]	10	11	24 [†]	23 [†]	33 [†]	30 [†]	
Exemplar	48 [†]	45 [†]	71 [†]	62 [†]	77 [†]	67 [†]	68 [†]	64 [†]	40 [†]	42 [†]	46 [†]	45 [†]	58 [†]	54 [†]	
location															
Urban	8 [*]	—	9 [‡]	—	5	—	6	—	1	—	2	—	—	—	5
Respondent role															
Behaviorist	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Manager	2	—	-9 [*]	—	-6	—	0	—	3	—	-1	—	—2	—	—2
Physician	-2	—	-3	—	-14 [‡]	—	-2	—	1	—	1	—	—3	—	—3
Practice size															
10+ employees	2	—	-3	—	13 [*]	6	—	-3	—	-5	2	—	—	—	—
Constant	31 [†]	27 [†]	23 [†]	29 [†]	21 [†]	15	18 [†]	11	43 [†]	43 [†]	27 [†]	31 [†]	27 [†]	26 [†]	

Note. In each domain, Model A is a linear regression of the domain score as a function of the level of integration alone. Model B also includes three potential confounders. Non-BHC = Practice with no Behavioral Health Clinicians; CMHC = Community Mental Health Center. Urban location includes inner city practices. Behaviorist includes student interns.

[†] $p < .001$. [‡] $p < .01$. ^{*} $p < .05$.

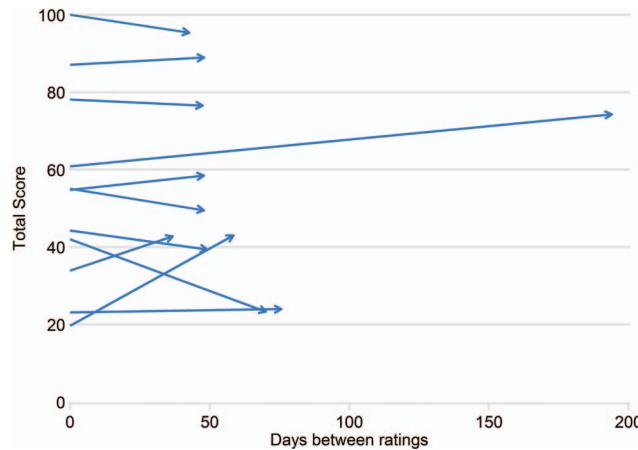


Figure 3. Intrarater consistency over time (test-retest reliability). Each arrow runs from a single respondent's initial score to their repeat score. See the online article for the color version of this figure.

Integration score among the 32 respondents was 7.1 with a range from 0 to 18. There was somewhat less agreement among respondents from the same practice in the other domains (see Table 4).

Discrimination Among Scenarios

Five PCPs or BHCs with integrated BH experience each completed the PIP for four written scenarios representing a range of practice settings. They were also asked to rank the four scenarios in terms of their overall degree of integration. The correlation between their rankings and their Total Integration Scores was significant (Spearman's $\rho = -0.71$; $P = 0.0005$). Correlations of their

overall rankings with their domain scores were Workflow $\rho = -0.58$ ($P = 0.007$), Clinical Services $\rho = -0.40$ ($P = 0.08$), Workspace $\rho = -0.89$ ($p < .0001$), Shared Care Plans $\rho = -0.67$ ($p = .002$), Case Identification $\rho = -0.47$ ($p = .04$), Patient Engagement $\rho = -0.24$ ($p = .33$).

Discussion

Validity

A test or instrument is valid for a particular purpose if it measures the underlying phenomenon or construct that it purports to measure and

Table 4
Reliability by Domain

Domain	Intrarater consistency over time ($N = 11$ subjects)			Interrater agreement within practice ($N = 32$ respondents from 15 practices)			
	Mean change	SD	95% limits of agreement	Mean diff	Min	Max	SD
Workflow	+2.7	16.7	-31, +36	12.8	0	38	8.6
Services	+2.3	16.3	-30, +35	11.2	0	36	7.2
Workspace	-5.7	15.2	-36, +25	14.5	0	50	9.5
Shared care	-5.1	16.0	-37, +27	11.7	0	25	7.3
Identification	+10.9	18.1	-25, +47	17.0	0	50	11.3
Engagement	+4.0	12.6	-21, +29	13.3	0	31	8.8
Total integration	+1.5	11.1	-21, +24	7.1	0	18	4.7

Note. SD = standard deviation; Diff = difference; Min = minimum difference; Max = maximum difference.

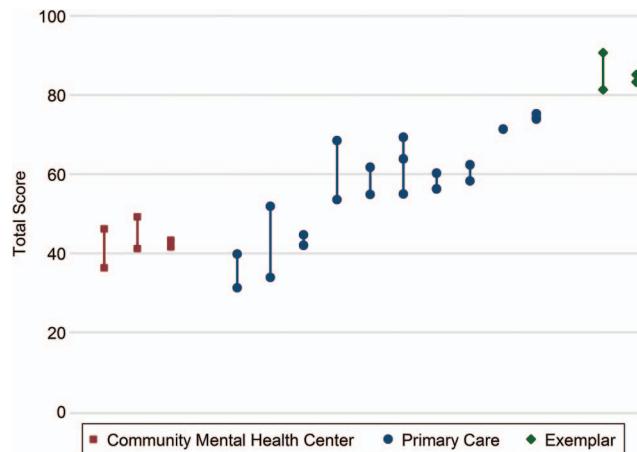


Figure 4. Interrater agreement within the same practice. Each vertical line represents one practice with two or three respondents. The points indicate the Total Integration Scores provided by each respondent. See the online article for the color version of this figure.

supports the conclusions that are drawn from it (McDowell, 2006). Validity has many components, including reliability, content and construct validity, and the ability to discriminate among phenomena that are importantly different.

Reliability

We measured reliability three ways. First, the internal consistency of the instrument as measured by Cronbach's alpha is quite high ($\alpha = 0.95$ for the Total Integration Score; see Table 4). Second, 11 respondents with repeat participation showed very little change in their responses (see Figure 3). Notably, there was no relationship between the amount of change and the time between the responses, which would have suggested that respondents simply remembered their previous answers. Finally, different respondents assessing the same practices showed a high level of agreement (see Table 4).

Content Validity

Content validity assesses whether the items chosen represent the underlying concepts or theoretical domain they are meant to reflect (Aday, 1996). In the case of the PIP, those domains are specified by the *Lexicon for Behavioral Health and Primary Care Integration* (Peek & the National Integration Academy Council, 2013). The PIP includes items representing all the *Lexicon*

domains, albeit sometimes combined with another related domain.

Construct Validity

Construct validity measures how well an instrument reflects the underlying target construct (in this case “integration of behavioral health and primary care”) to the exclusion of other characteristics. In the absence of a gold-standard reference test for integration, we rely on findings such as the PIP’s ability to discriminate among practices with *prima facie* differences in integration in both real-world settings (see Table 2) and artificial scenarios. Construct validity is further supported by the observation that the PIP’s ability to discriminate is not confounded by the practice location or size or the role of the respondent (see Table 3).

Strength, Weaknesses, Limitations and Future Directions

In the absence of a “gold standard” test for integration, it is impossible to determine the criterion validity (sensitivity and specificity) of the PIP. However, the use of the four levels of integration as a reference point increase our confidence that the PIP scores represent what observers of the field mean by “Integrated Behavioral Health.”

Although the respondents were a convenience sample, they derive from a broad range of practices across many settings and in various stages of integration, suggesting that they may generalize well to other settings where the PIP is intended for use. Although the respondents included a broad range of raters from US practices, we have no information about PIP performance outside the US.

Because we requested repeat measures from only a small number of raters, conclusions about the within-rater reliability of the PIP are limited by small sample size.

The PIP is a measure of the structures and processes in place and does not record patient outcomes, financial performance, population health or other desired aspects of high quality care. Nonetheless, structure and process are two of the three essential aspects of quality (Donabedian, 1988) and must be measured to allow thoughtful and effective management.

Experience with the PIP is still relatively small. As more practices and researchers use it for quality management, identification of best practices, process redesign, assessment of interventions, and other health services analyses, we will learn more about its strengths and limitations. Additional opportunities remain to improve the items and apply the results of the PIP in other countries and languages. A version is being planned for use in China. The wording of items can be further improved to increase the measure's reliability. The creation of a companion measure that can be completed by patients has potential to enhance the PIP's validity.

Conclusion

Initial experience with the PIP suggests good feasibility and face validity, low response burden, high within-subject reliability, and good discrimination.

References

Aday, L. A. (1996). *Designing and conducting health surveys: A comprehensive guide*. San Francisco, CA: Jossey-Bass.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal (Clinical Research Ed.)*, 314, 572. <http://dx.doi.org/10.1136/bmj.314.7080.572>

Blount, F. A., & Miller, B. F. (2009). Addressing the workforce crisis in integrated primary care. *Journal of Clinical Psychology in Medical Settings*, 16, 113–119. <http://dx.doi.org/10.1007/s10880-008-9142-7>

Cohen, D. J., Balasubramanian, B. A., Davis, M., Hall, J., Gunn, R., Stange, K. C., . . . Miller, B. F. (2015). Understanding care integration from the ground up: Five organizing constructs that shape integrated practices. *Journal of the American Board of Family Medicine*, 28 (Suppl. 1), S7–S20. <http://dx.doi.org/10.3122/jabfm.2015.S1.150050>

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. <http://dx.doi.org/10.1007/BF02310555>

Cuzick, J. (1985). A Wilcoxon-type test for trend. *Statistics in Medicine*, 4, 87–90. <http://dx.doi.org/10.1002/sim.4780040112>

Dickinson, W. P. (2015). Strategies to Support the integration of behavioral health and primary care: What have we learned thus far? *Journal of the American Board of Family Medicine*, 28, S102–S106. <http://dx.doi.org/10.3122/jabfm.2015.S1.150112>

Donabedian, A. (1988). The quality of care. How can it be assessed? *Journal of the American Medical Association*, 260, 1743–1748. <http://dx.doi.org/10.1001/jama.1988.03410120089033>

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, 42, 377–381. <http://dx.doi.org/10.1016/j.jbi.2008.08.010>

Hunter, C. L., Goodie, J. L., Dobmeyer, A. C., & Dorrance, K. A. (2014). Tipping points in the Department of Defense's experience with psychologists in primary care. *American Psychologist*, 69, 388–398. <http://dx.doi.org/10.1037/a0035806>

Institute for Healthcare Improvement. (2015). *IHI Triple Aim Measures*. Retrieved from <http://www.ihi.org/engage/initiatives/TripleAim/Pages/MeasuresResults.aspx>

Kessler, R., van Eeghan, C., Mullin, D., Auxier, A., Macchi, C. R., & Littenberg, B. (2015). Measuring behavioral health integration in primary care settings. *The Health Psychologist*. [Manuscript submitted for publication].

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.). Oxford, UK: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780195165678.001.0001>

Miller, B. F. (2015). When frontline practice innovations are ahead of the health policy community: The example of behavioral health and primary care integration. *Journal of the American Board of*

Family Medicine, 28, S98–S101. <http://dx.doi.org/10.3122/jabfm.2015.S1.150107>

Peek, C. J., & the National Integration Academy Council. (2013). *Lexicon for Behavioral Health and Primary Care Integration* (AHRQ Publication No. 13-IP001-EF). Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from <http://integrationacademy.ahrq.gov/sites/default/files/Lexicon.pdf>

Spearman, C. E. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15, 72–101.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley Pub. Co.

Appendix

The Practice Integration Profile (www.uvm.edu/~pip/pip.php)

Instructions: We suggest that it be rated both by the medical director and a senior behavioral health clinician. First, please check that you have reviewed the terms and conditions. Then, read the statements in each of the eight dimensions and select the response that best reflects your organization. Most items ask for a rough approximation of how often your practice meets a particular criterion and with a numerator and denominator to guide your thinking. You don't need to collect specific data - just provide your best estimate. Where we refer to "patients," feel free to consider family, caregivers, surrogates and other stakeholders as appropriate. Some items are ordered such that each level implies that all the previous criteria are met. Please choose the highest level that applies based on current practice activities.

In our practice, . . .	Examples	Scoring criteria	Score				
Practice workflow (PW)							
WF1 . . . we use a standard protocol for patients who need or can benefit from integrated Behavioral Health (BH).	Patients in need of BH services are identified, assessed and receive care using a consistent set of processes	Numerator = # of patients receiving protocol-based care Denominator = # of patients in need of BH	Never	Sometimes	Often	Frequently	Always
WF2 . . . we use registry tracking for patients with identified BH issues.	Insomnia or depression registry	Numerator = # of patients in BH registries Denominator = # of patients with BH needs	Never	Sometimes	Often	Frequently	Always
WF3 . . . we provide coordination of care for patients with identified BH issues.	We coordinate appointments with outside medical and non-medical providers, or assist with social services contacts	Numerator = # of patients receiving coordinated care Denominator = # of patients with BH needs	Never	Sometimes	Often	Frequently	Always

(Appendix continues)

In our practice, . . .	Examples	Scoring criteria	Score					
			Never	Sometimes	Often	Frequently	Always	
WF4	. . . we provide referral assistance to connect patients to community resources,	Exercise programs, AA, housing assistance, support groups, etc.	Numerator = # of patients receiving referral assistance to community resources Denominator = # of patients needing referral to community resources	0% 0%	1–33% 1–33%	34–66% 34–66%	67–99% 67–99%	100% 100%
WF5	. . . we provide referral assistance to connect patients to specialty mental health resources.	Psychiatry for persistent severe mental illness	Numerator = # of patients receiving referral assistance to specialty mental health resources Denominator = # of patients needing referral to specialty mental health resources	Never 0%	Sometimes 1–33%	Often 34–66%	Frequently 67–99%	Always 100%
WF6	. . . we use a standard approach for documenting patients' self-management goals.	Goals are documented in a structured problem list or other well-defined place	Numerator = # of patients with documented goals Denominator = # of patients with BH needs	Never 0%	Sometimes 1–33%	Often 34–66%	Frequently 67–99%	Always 100%
Clinical services (CS)								
CS1	. . . we have clinicians available on site who provide non-crisis focused BH services.	Scheduled care (assessment, counseling, referral, etc.) of behavioral issues	Numerator = # hours non-crisis BH services are available Denominator = # of hours the clinic is open	Never 0%	Sometimes 1–33%	Often 34–66%	Frequently 67–99%	Always 100%
CS2	. . . we have clinicians available on site to respond to patients in behavioral crisis.	Urgent care of patients in behavioral crisis	Numerator = # hours crisis BH services are available Denominator = # of hours the clinic is open	Never 0%	Sometimes 1–33%	Often 34–66%	Frequently 67–99%	Always 100%
CS3	. . . we have BH clinicians who can respond to seriously mentally ill and substance-dependent patients.	Schizophrenia, problem drinking, etc.	Numerator = # hours BH services for seriously mentally ill and substance-dependent patients are available Denominator = # of hours the clinic is open	Never 0%	Sometimes 1–33%	Often 34–66%	Frequently 67–99%	Always 100%
CS4	. . . we offer behavioral interventions for patients with chronic/complex medical illnesses.	Assessment, counseling, coaching for BH needs of diabetes, cancer, heart disease, hypertension, etc.	Numerator = # of patients offered BH interventions for chronic/complex medical illnesses Denominator = # of patients needing such services	Never 0%	Sometimes 1–33%	Often 34–66%	Frequently 67–99%	Always 100%

(Appendix continues)

In our practice, . . .	Examples	Scoring criteria	Score				
			Never	Sometimes	Often	Frequently	Always
CS5	. . . we employ BH clinicians with a background and training in complex or specialized behavioral health therapies.	Numerator = # of BH staff with training in complex or specialized behavioral health therapies Denominator = # of BH staff	0%	1–33%	34–66%	67–99%	100%
CS6	. . . we offer evidence-based substance abuse interventions.	Screening and brief intervention	Numerator = # of patients offered evidence-based substance abuse interventions Denominator = # of patients needing such services	0%	1–33%	34–66%	67–99%
CS7	. . . we offer prescription medications for routine mental health and substance abuse diagnoses.	Moderate depression and anxiety	Numerator = # of patients offered prescription medications for routine mental health or substance abuse diagnoses Denominator = # of patients needing such services	0%	1–33%	34–66%	67–99%
CS8	. . . we offer prescription medications for serious complex co-occurring mental health and/or substance abuse diagnoses	Major depression, bi-polar, schizophrenia	Numerator = # of patients offered prescription medications for serious mental health or substance abuse diagnoses Denominator = # of patients needing such services	0%	1–33%	34–66%	67–99%
CS9	. . . we offer referral to non-clinical services outside of our practice.	Spiritual advisors, schools, criminal justice (probation and parole, drug courts), or vocational rehabilitation	Numerator = # of patients offered referrals Denominator = # of patients needing such services	0%	1–33%	34–66%	67–99%
Workspace (WS)							
WS1	. . . BH and Medical Clinicians work in:	Shared building or unit	Ordered—Please pick the best descriptor of your practice	Different Buildings	Different Floors	Different Office Suites	Separate parts of the same suite Fully shared space
WS2	. . . patient treatment/ care plans are routinely documented in a medical record accessible to both BH and medical clinicians.	Medical and BH clinicians use the same Electronic Record	Numerator = # of patients with shared records Denominator = # of patients receiving BH services	Never	Sometimes	Often	Frequently Always

(Appendix continues)

In our practice,	Examples	Scoring criteria	Score				
Integration and sharing methods (IN)							
IN1	... BH and Medical Clinicians regularly and actively exchange information about patient care.	Active includes "tasking" or both clinicians signing shared documentation. Does not include simply documenting in a place that is available to the other clinician	Numerator = # of patients with regular active exchange of information	Never	Sometimes	Often	Frequently
			Denominator = # of patients receiving BH services	0%	1-33%	34-66%	67-99%
							100%
IN2	... there are regular educational activities including both BH and Medical Clinicians	This includes but is not limited to sessions focused on specific conditions such as patients with chronic pain or depression. Includes case conferences, seminars, etc.	Educational activities should be jointly provided to medical and behavioral clinicians.	No structured educational activities	Educational activities are provided to BH and medical clinicians separately	Some activities with both medical and BH clinicians	Frequent activities with both medical and BH clinicians
							Regularly scheduled activities with full participation by both medical and BH clinicians
IN3	... BH and Medical Clinicians regularly spend time together collaborating on patient care.	Face-to-face contact to discuss patient care	Numerator = # of patients discussed in person	Never	Sometimes	Often	Frequently
			Denominator = # of patients receiving BH services	0%	1-33%	34-66%	67-99%
							100%
IN4	... patients with BH needs have shared care plans developed jointly by the patient, BH and Medical clinicians.	Joint visits with patient, caregivers, medical and BH clinicians for development of a problem list and action plan; iterative development of the problem list and plan by individual clinicians meeting with the patient/ caregivers.	Numerator = # of patients with a shared care plan	Never	Sometimes	Often	Frequently
			Denominator = # of patients receiving BH services	0%	1-33%	34-66%	67-99%
							100%
Case identification (ID)							
ID1	... we screen eligible adults for BH conditions using a standardized procedure.	US Preventative Services Task Force guidelines for alcohol use, depression, etc.	Numerator = # screened	Never	Sometimes	Often	Frequently
			Denominator = # of adults seen in the practice	0%	1-33%	34-66%	67-99%
							100%
ID2	... we use practice-level data to screen for patients at risk for complex or special needs.	Billing, registration data, disease registry, lab results, etc.	Numerator = # of patients screened	Never	Sometimes	Often	Frequently
							Always

(Appendix continues)

In our practice, ...		Examples	Scoring criteria	Score				
			Denominator = # of patients in the practice	0%	1-33%	34-66%	67-99%	100%
ID3	... patients are screened at least annually for behavioral conditions related to a medical problem.	Screening for depression in diabetes, anxiety in heart failure, etc.	Numerator = # screened	Never	Sometimes	Often	Frequently	Always
ID4	... all patients are screened at least annually for lifestyle or behavioral risk factors	Poor diet, inadequate exercise, sleep disorders, etc.	Denominator = # of patients with target medical conditions	0%	1-33%	34-66%	67-99%	100%
ID5	... screening data are presented to clinicians with recommendations for patient care.	Patients with low physical activity are flagged for physician to consider referral to YMCA; patients with insomnia are flagged for referral to CBT.	Numerator = # of recommendations presented to clinician	Never	Sometimes	Often	Frequently	Always
			Denominator = # of positive findings (patients with multiple positive screens are counted multiple times)	0%	1-33%	34-66%	67-99%	100%
Patient engagement (PE)								
PE1	... we successfully engage identified patients in Behavioral Care	Patients who need counseling actually start counseling	Numerator= # initiating behavioral intervention	Never	Sometimes	Often	Frequently	Always
			Denominator = # of patients who are identified with a specific behavioral need	0%	1-33%	34-66%	67-99%	100%
PE2	... we successfully retain patients in Behavioral Care	Patients who initiate counseling complete counseling	Numerator= # completing behavioral intervention	Never	Sometimes	Often	Frequently	Always
			Denominator = # of patients who initiate behavioral intervention	0%	1-33%	34-66%	67-99%	100%
PE3	... have specific systems to identify and intervene on patients who did not initiate or complete care	Post-referral "tickler" files with staff follow-up	Numerator = # receiving action to engage or retain	Never	Sometimes	Often	Frequently	Always
			Denominator = # of patients who do not initiate or complete BH care	0%	1-33%	34-66%	67-99%	100%
PE4	... we have follow-up plans for all patients who complete BH interventions	Automatically scheduled visits with primary care provider	Numerator = # of patients with a specific follow-up plan	Never	Sometimes	Often	Frequently	Always

In our practice, ... Examples	Scoring criteria	Score				
	Denominator = # of patients who complete a BH intervention	0%	1–33%	34–66%	67–99%	100%

Received April 18, 2016
Revision received July 21, 2016
Accepted July 25, 2016 ■