



HHS Public Access

Author manuscript

Psychiatr Serv. Author manuscript; available in PMC 2018 September 01.

Published in final edited form as:

Psychiatr Serv. 2017 September 01; 68(9): 975–978. doi:10.1176/appi.ps.201600264.

Is There a Role for Fidelity Self-Assessment for IPS?

Paul J. Margolies,

Columbia Univ - New York State Psychiatric Institute, New York, New York

Jennifer L. Humensky,

Center of Excellence in Cultural Competence, New York State Psychiatric Institute, New York, New York

I-Chin Chiang,

New York State Psychiatric Institute, Columbia University, New York, New York

Nancy H. Covell,

Mental Health Services and Policy Research, New York State Psychiatric Institute, New York, New York

Karen Broadway-Wilson,

New York State Psychiatric Institute, Columbia University, New York, New York

Raymond Gregory,

New York State Psychiatric Institute, Columbia University, New York, New York

Thomas Jewell,

New York State Psychiatric Institute, Columbia University, New York, New York

Gary Scannevin,

New York State Psychiatric Institute, Columbia University, New York, New York

Stephen Baker, and

IPS Consultant, Washington, D.C

Lisa B. Dixon

New York State Psychiatric Institute, and Dept of Psychiatry, Columbia University, New York, New York

Abstract

Objective—Fidelity assessments help ensure evidence-based practices are implemented properly. While typically conducted by independent raters, some programs have implemented self-assessments due to resource constraints. Self-assessments were compared to independent assessments within programs implementing Individual Placement and Support (IPS) supported employment.

Methods—Eleven community-based outpatient programs in New York State completed both self and independent assessments. ICCs and paired t-tests were used to compare independent and self-rated assessments.

The authors have no conflicts of interest to report.

Results—Mean scores are within the range of “fair fidelity” to IPS. Mean self (91.7) and independent (92.9) scores were not significantly different from one another. However, significant differences were found among individual items in this small sample.

Conclusions and Implications for Practice—Self-assessments may be valid for examining a program’s overall functioning, and useful when resource constraints prevent independent assessment. Independent assessors may be able to identify nuances, particularly among individual assessment items, that can identify areas for program improvement.

The growth of evidence-based practices (EBPs) in mental health treatment, particularly since the late 1990s (1), has increased the demand for fidelity assessment (2). EBPs have been demonstrated to be effective and are expected to achieve similar results when implemented in new treatment settings (2). Supported employment (SE) as operationalized by the Individual Placement and Support (IPS) model is one EBP that has demonstrated effectiveness in improving vocational outcomes for persons with mental health disorders (3,4). However, when implemented in a new site, with new personnel, the model may not be implemented properly and thus may not achieve the same results (5).

Fidelity scales examine the extent to which a program is implementing core principles and procedures of the EBP (6). Assessors follow a protocol to gather information from a variety of sources. In-person visits typically include interviews with multiple stakeholders, including program leadership, staff implementing the program, and clients. Program documentation including client charts and other clinical records are typically reviewed (2).

Independent fidelity assessment can be expensive and time-consuming, and as the number of EBPs grows, it can be burdensome for agencies to identify qualified assessors. The intensive one to two day process can be burdensome for program sites (7). Consequently, some programs have begun conducting self-assessments to complement and supplement independent assessments (7), for example, doing self- and independent assessments in alternate years. Studies of Assertive Community Treatment (ACT), have shown that self- and independent assessments can yield comparable results under some circumstances (8,9). However, these results may not be generalizable to all EBPs; self-assessments may be best in stable programs with a history of good fidelity (8), who are following a defined protocol (7).

We examine how assessment methods compare within an IPS program model where programs receive extensive training and support to collect self-reported data following the IPS fidelity protocol.

Methods

Fidelity assessments were conducted by program staff (“self-assessments”) and independent expert raters (“independent assessments”) at 11 Personalized Recovery Oriented Services (PROS) programs across New York State (NYS). PROS is an outpatient mental health program model that sets a clear expectation in regard to the implementation of recovery-oriented evidence-based practices. Through funding policies, the NYS Office of Mental Health provides incentives for adoption of these practices, which include IPS (10).

Fidelity assessments were one component of a comprehensive training and implementation technical assistance package offered to PROS programs across NYS by the Center for Practice Innovations (CPI) (10). Programs participated in regional learning collaboratives providing face-to-face and online training and support.

Continuous quality improvement process served as the foundation for learning collaborative activity; participating programs routinely collecting and sharing data including performance indicators and fidelity ratings. Learning collaborative leaders structured the process so that programs experienced the use of data as helpful for their implementation efforts and not punitive. In the learning collaboratives, PROS program staff were taught about IPS fidelity generally, and how to conduct fidelity self-assessments specifically, through webinars and program-specific consultation calls and visits.

A total of 52 PROS programs completed fidelity self-assessments during the last quarter of 2014. Programs used the IPS Supported Employment Fidelity Scale (3, 11), which consists of 25 items clustered into three sections (staffing, organization, and services). Each item consists of a 5 point scale and the maximum total score is 125.

The programs completing self-assessments were clustered into four regions and, within each region, randomized. Within each region, programs were contacted following the order of randomization, and asked to voluntarily participate in an independent fidelity assessment. A total of 20 programs were contacted before 3 programs in each region agreed to participate. Of these 12 programs, one was not assessed due to scheduling issues. These independent ratings occurred during the second quarter of 2015. The time between the 2014 self-assessments used in this analysis and independent assessments ranged from two to eight months, with a mean of five months. The 8 invited, but non-participating programs cited lack of time, lack of interest, or simply did not respond to requests. Mean scores between the 11 participating programs self-assessments and the 8 refusing programs were not statistically different from one another. They were also not statistically different from the 41 other programs assessed.

Two independent raters, external to the agencies and to CPI, conducted the independent assessments. One rater was trained by the developers of IPS and has conducted independent assessments for many years. The other rater was trained by the first rater through didactics, modeling and coaching. Two independent assessments were conducted by both raters and 9 were conducted by one of the two raters. The number of interviews varied by the composition of program staff, but generally included the program director, supported employment supervisor, one or more supported employment workers, one or more clinicians, and up to five clients. In addition, assessors reviewed clinical documentation including a sample of client charts, supported employment caseload, and job development logs. The independent assessments were completed in one day due to the typically small scale of IPS implementation at these program sites (only 2 of the 11 programs had more than 1.0 FTE employment staff). For comparison, among the 130 programs participating in the IPS Learning Community nationwide, the median was three IPS specialists per program (Gary Bond, personal communication, October 31, 2016).

Fidelity scores were compared between independent and self-assessments, using paired t-tests and two-way mixed effects intraclass correlations (ICC) with consistency of agreement (individual measurement). We also examined the effect size of the differences between the assessments, using Cohen's d. Analyses were conducted using IBM SPSS Statistics version 23. This program evaluation did not constitute human subjects research as defined by the Institutional Review Board of the New York State Psychiatric Institute.

Results

Mean total scores (Table 1) between self and independent assessments were not significantly different from one another, and indicated fair inter-rater agreement (ICC of 0.52) (12). These scores are within the range (scores of 75–99 out of a total of 125) of IPS guidelines for “fair fidelity” to the IPS model (11). Independent assessments found 3 programs with good fidelity (total scores >99), 7 with “fair” fidelity (total scores 75–99), and 1 “not IPS” (total score <75). Self assessments found 4 programs with good fidelity (total scores >99), 6 with “fair” fidelity (total scores 75–99), and 1 “not IPS” (total score <75).

While the mean scores did not differ significantly, we did find significant variation among some of the individual items. Two items showed significant differences between the self and independent assessments in paired t-tests: time unlimited follow-along supports ($p=.01$) and work incentives planning ($p=.04$). Additionally, seven of the 25 items had differences between assessments that approach a medium effect size (at or above Cohen's d of .4). Moreover, ICCs on eight of the 25 items were below zero, which can occur in two-way mixed-effects ICC models, and another five had ICCs below .4, indicating poor interrater agreement (12). Thus, some variability in individual items was observed, albeit in this small sample size.

Discussion

Is there a place for fidelity self-assessments? This issue has received attention recently (7, 13) and, given demand for increased fidelity assessment with widespread adoption of evidence-based practices, will continue to benefit from close examination. Bond (13) thoughtfully cautions against replacing independent fidelity reviews with self-assessments, while at the same time noting the usefulness of self-assessment for quality improvement. Can self-assessments be trusted? If so, under what conditions? The data presented here may help move the discussion along.

Across the 11 programs, there were no statistical differences between total fidelity score means which were within the range of “fair fidelity”. This suboptimal fidelity points to opportunities across the state and within individual programs for continuous quality improvement efforts. Only two items were significantly different between the assessments. Independent raters gave a lower rating (average of .91 point difference) to estimates of time-unlimited follow-along supports. This had a complicated definition within the PROS programs, as persons step down from intensive PROS services to less-intensive Ongoing Rehabilitation and Support services once they obtain a competitive job. Thus, it is possible that the programs and external assessors interpreted continuity of care between intensive and

stepped-down services differently. Work incentives planning was rated higher by independent than self-assessors (average difference .73 points), which may reflect modesty by programs in incentives planning, changes in programs between the assessment times, or other differences in interpretation. We also found some variability across items, as measured by low ICCs and large Cohen's *d* effect size differences, albeit in this small sample of 11 programs. If this variation is found to be stable across other samples, it may indicate that, in some cases, self assessments may provide a valid snapshot of overall program functioning, but that independent assessors may be better at identifying nuanced areas for improvement in individual items.

Given biases often found with self-report (14, 15), several conditions may have contributed to these findings. This fidelity scale is well designed and contains many concrete details and operational definitions to guide its use. This "user friendly" aspect should not be overlooked. As noted previously, PROS program staff were taught about IPS fidelity and how to conduct fidelity self-assessments. It appears as if they learned this well. It is also possible that the learning collaboratives' emphasis on continuous quality improvement resulted in an implementation environment that was experienced as safe enough for participants to report data honestly and without bias. Though this is speculation, our ongoing contact with, and knowledge about, programs may result in less likelihood of dishonest reporting.

This study has clear limitations including limited sample size, five month average between the two methods of assessment, small number of employment staff per program, significant amount of training made available to program staff (which may not be representative of training typically available to those attempting to use self-assessment), and inability to empirically test the conditions contributing to the findings. Future studies may choose to address these issues as well as attempt to answer important questions such as when fidelity self-assessments may (and may not) be appropriate, what circumstances indicate the need for independent assessors, and, when used for continuous quality improvement, whether there is a difference in the impact of self-assessments vs. independent assessments.

Conclusions

This study, using the IPS Supported Employment Fidelity Scale, focused on the relationship between fidelity self-assessment and independent assessment. There were no significant differences between the total fidelity score means (i.e., self assessments vs. independent assessments) across 11 community mental health programs. However, we found some variation in individual items; future research should examine whether these trends continue in larger samples. These results may suggest that self-assessments may be useful under certain circumstances, but that independent assessors may be able to identify nuances and differences in individual items. Both self and independent assessments may be useful to programs and policymakers in appropriate contexts.

Acknowledgments

Dr. Humensky receives salary support from the National Institute of Mental Health (grant number K01MH103445).

References

1. Aarons GA, Farahnak LR, Ehrhart MG, et al. Aligning Leadership Across Systems and Organizations to Develop Strategic Climate to for Evidence-Based Practice Implementation. *Annual Review of Public Health*. 2014; 35:255.
2. Bond GR, Drake RE, McHugo GJ, et al. Strategies for improving fidelity in the national evidence-based practices project. *Research on Social Work Practice*. 2009; 19:569–581.
3. The IPS Employment Center, Rockville Institute. <https://www.ipsworks.org/Fidelity> assessment available at <https://www.ipsworks.org/wp-content/uploads/2014/04/IPS-Fidelity-Scale-Eng1.pdf>
4. Bond GR, Drake RE. Making the case for IPS supported employment. *Administration and Policy in Mental Health and Mental Health Services Research*. 2014; 41:69–73. [PubMed: 23161326]
5. Hurlburt M, Aarons GA, Fettes D, et al. Interagency collaborative team model for capacity building to scale-up evidence-based practice. *Children and Youth Services Review*. 2014; 39:160–168. [PubMed: 27512239]
6. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*. 2011; 38:65–76. [PubMed: 20957426]
7. McGrew J, White L, Stull L. Self-assessed fidelity: proceed with caution: In Reply. *Psychiatric Services*. 2013; 64:394.
8. McGrew JH, White LM, Stull LG, et al. A comparison of self-reported and phone-administered methods of ACT fidelity assessment: a pilot study in Indiana. *Psychiatric Services*. 2013; 64:272–276. [PubMed: 23280337]
9. Rollins AL, McGrew JH, Kukla M, et al. Comparison of Assertive Community Treatment Fidelity Assessment Methods: Reliability and Validity. *Administration and Policy in Mental Health and Mental Health Services Research*. 2016; 43:157–167. [PubMed: 25721146]
10. Margolies PJ, Broadway-Wilson K, Gregory R, et al. Use of learning collaboratives by the Center for Practice Innovations to bring IPS to scale in New York State. *Psychiatric Services*. 2015; 66(1): 4–6. [PubMed: 25321348]
11. Luciano A, Bond GR, Drake RE, et al. Is high fidelity to supported employment equally attainable in small and large communities? *Community Mental Health Journal*. 2014; 50:46–50. [PubMed: 24352684]
12. Cicchetti DV. Guidelines, criteria and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 1994; 6:284–290.
13. Bond GR. Self-assessed fidelity: proceed with caution. *Psychiatric Services*. 2013; 64:393–394.
14. He J, van de Vijver FJ. Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences*. 2015; 81:129–134.
15. McGrath RE, Mitchell M, Kim, et al. Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*. 2010; 136:450. [PubMed: 20438146]

Table 1 Differences between IPS Independent and Self-Assessments, by Item [range highest to lowest in Independent Assessment]

Item	Independent Assessment		Self-Assessment		Mean Difference	p-value	ICC	Cohen's d
	Mean	SD	Mean	SD				
Total ^a	92.9	10.77	91.7	13.74	1.2	.75	.52	.10
SE2. Disclosure ^b	4.91	.30	4.73	.65	.18	.17	.68	.37
O2. Integration of rehabilitation with mental health treatment through frequent team contact	4.73	.47	4.36	1.03	.36	.22	.33	.47
O6. Zero Exclusion Criteria	4.73	.90	4.82	.40	-.09	.78	-.11	-.14
SE3. Ongoing, work-based, vocational assessment	4.73	.47	4.18	1.25	.55	.24	-.16	.61
S1. Caseload Size	4.64	.92	4.18	1.40	.45	.27	.41	.40
SE5. Individualized job search	4.64	.81	4.55	.69	.09	.59	.74	.13
SE10. Competitive Jobs	4.27	1.62	3.91	1.30	.36	.61	-.22	.26
SE9. Diversity of employers	4.18	1.47	4.09	1.14	.09	.80	.63	.07
S3. Vocational Generalists	4.18	.87	3.91	1.22	.27	.57	-.07	.27
O7. Agency focus on competitive employment	4.18	.87	4.45	.93	-.27	.59	-.60	-.31
SE1. Work Incentives Planning	4.18	1.08	3.45	1.29	.73	.04*	.64	.64
SE11. Individualized follow-along supports	4.00	1.73	4.18	1.25	-.18	.55	.79	-.13
SE4. Rapid search for competitive job	3.91	1.22	3.82	1.25	.09	.88	-.21	.08
SE14. Assertive engagement and outreach by integrated team	3.82	.75	3.09	1.45	.73	.09	.39	.66
SE7. Job development: quality of employer contacts	3.73	1.42	3.55	1.44	.18	.51	.81	.13
SE12. Follow-along supports: time unlimited	3.55	1.13	4.45	.69	-.91	.01*	.49	-1.02
S2. Vocational Services Staff	3.45	1.57	2.91	1.45	.55	.34	.28	.38
O1. Integration of rehabilitation with mental health treatment through team assignment	3.36	1.75	3.91	1.30	-.55	.35	.27	-.37
SE8. Diversity of jobs developed	3.36	1.63	4.18	1.17	-.82	.22	-.09	-.61
SE6. Job development: frequent employer contact	2.82	1.83	3.09	1.45	-.27	.54	.63	-.17
O5. Role of Employment Supervisor	2.64	1.29	2.82	1.40	-.18	.69	.40	-.14
O8. Executive Team Support for SE	2.45	1.63	2.91	1.14	-.45	.50	-.18	-.34
O3. Collaboration between employment specialists and Vocational Rehabilitation	2.18	1.83	1.73	1.10	.45	.14	.81	.31
SE13: Community Based Services	2.18	1.17	2.27	1.27	-.09	.80	.57	-.08

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Item	Independent Assessment		Self-Assessment		Mean Difference	p-value	ICC	Cohen's d
	Mean	SD	Mean	SD				
O4: Vocational unit	2.09	1.87	2.18	1.33	-.09	.87	.37	-.06

* p<(0.05)

^a Possible total fidelity scores range from 25 to 125, with higher scores indicating greater fidelity to the IPS model.

^b Possible scores on each of the 25 composite items range from 1 to 5, with higher scores indicating full implementation of that item.