Published in final edited form as:

Psychiatr Serv. 2013 March 1; 64(3): 272–276. doi:10.1176/appi.ps.001252012.

A comparison of self-reported and phone-based fidelity for Assertive Community Treatment (ACT): A pilot study in Indiana

John H. McGrew, Ph.D.,

Department of Psychology at Indiana University–Purdue University Indianapolis

Laura M. White, M.S.,

Department of Psychology at Indiana University–Purdue University Indianapolis

Laura G. Stull, Ph.D., and

Department of Psychology at Indiana University–Purdue University Indianapolis

Ms. Jennifer Wright-Berryman, MSW

Adult and Child Mental Health Center, Indianapolis

Abstract

Objective—The time and effort associated with monitoring fidelity of evidence-based practices is costly. This study investigated the reliability and validity of a less burdensome approach: self-reported fidelity.

Methods—Phone-based and self-reported fidelity were compared for 16 ACT teams. Team leaders completed a self-report protocol providing information sufficient to score the Dartmouth Assertive Community Treatment Scale (DACTS). Two raters scored the DACTS based solely on the self-report protocol. Two additional raters conducted phone interviews with team leaders, verifying and updating the self-report protocol, and independently scored the DACTS.

Results—Self-report DACTS total scores were reliable and valid compared to phone assessment, based on interrater consistency (intraclass correlation) and consensus (mean rating differences). Phone assessment agreed with self-report within .25 scale points (out of 5) for 94% of sites.

Conclusions—Self-report fidelity could address concerns regarding costs of program monitoring as part of a stepped approach to quality assurance.

Although verification of program fidelity (1) is helpful in addressing the problem of inadequate implementation of evidence based practices (2) and the associated decrease in program outcomes (3,4), it is time intensive, often requiring a day or more for an onsite visit and another day to score and write the report. A 2007 national task force identified several innovative approaches to address practical concerns like costs and burden of quality improvement (5), including the use of alternative fidelity methods such as phone-administered assessments. In a prior study of Assertive Community Treatment, we found that phone fidelity assessment was reliable, and produced scores consistent with, and

comparable to, those from onsite assessment. To facilitate phone-administered fidelity, we created a self-report protocol (6) that required team leaders to gather and report information sufficient to score each fidelity item from the Dartmouth Assertive Community Treatment Scale (DACTS) (7). Our impression was that phone assessment mostly verified information already captured from the self-report protocol, leading us to speculate that self-report assessment might be an even less burdensome, alternative fidelity assessment method. In the current study, we examined the interrater reliability and concurrent validity of self-reported fidelity and phone-administered fidelity in a new data collection.

Methods

Twenty-four ACT teams in Indiana were invited to participate. Eight teams declined, having chosen not to maintain ACT certification due to changes in state funding. All 16 participating programs (66.7%) had been in operation for a minimum of 5 years, followed Indiana ACT standards, and received annual fidelity assessments to verify certification as an ACT provider (6). Data were collected between December 2010 and May 2011. ACT team leaders provided written informed consent to participate in the study. Study procedures were approved by the IUPUI Institutional Review Board.

The 28-item DACTS (7) was used to assess ACT fidelity. The DACTS provides a total score and three subscale scores: Human Resources (e.g., psychiatrist on staff), Organizational Boundaries (e.g., explicit admission criteria), and Nature of Services (e.g., in-vivo services). Items are rated using a 5-point behaviorally-anchored scale (5=fully implemented, 1=not implemented). Items with scores of 4 and higher are considered well-implemented. The DACTS has good interrater reliability (8) and can differentiate between ACT and other types of intensive case management (7).

A self-report fidelity protocol was used to collect all data needed to score the DACTS (6,7). The protocol consists of nine tables designed to summarize data efficiently: Staffing, Caseload/Discharges, Admissions, Hospitalizations, Client Contact Hours/Frequency, Services Received Outside of ACT, Engagement, Substance Abuse Treatment, and Miscellaneous (program meeting, practicing team leader, crisis services, and work with informal supports). A critical aspect of the protocol is the conversion of subjective, global questions into objective, focused questions, (e.g., rather than asking for a global evaluation of treatment responsibility, team leaders record the number of clients receiving a list of services outside ACT during the past month).

Sites received the self-report fidelity protocol two weeks before the phone interview. Team leaders consulted clinical and other program records to complete the protocol and returned it before the call. Sites contacted the research team with questions.

Phone interviews were conducted with the ACT team leader. In 3 (18.8%) cases, additional individuals participated in the call as an observer (e.g., medical director). Phone interviews were conducted jointly by the first and second authors and focused on reviewing the self-report fidelity protocol for accuracy. When the data were incomplete (10 teams), data either were identified by research staff and provided by the site prior to the call (2 cases), were

gathered during the phone interview (6 cases), or were submitted within one week after the call (2 cases). Raters independently updated the self-report fidelity protocol to reflect information gathered during or after the interview, and independently scored the DACTS based on the revised information. Discrepant DACTS items were then identified and raters met to discuss and assign the final consensus scores.

The self-report assessment was conducted by two new raters (third and fourth authors) who did not participate in the phone assessments. They independently scored the DACTS based solely on the self-report fidelity protocol as originally provided by the sites or, as amended with missing data provided before the phone call (2 sites). Self-report scoring was completed after the phone interviews, but did not include information obtained during the phone interviews. Raters left DACTS items blank (unscored) if data were unscorable/missing. Discrepant DACTS items were then identified and raters met to discuss and assign consensus scores. All four raters had at least one year of training/experience in conducting DACTS assessments as part of prior fidelity studies.

Two indicators were used to assess interrater (reliability) and intermethod (validity) agreement: consistency, calculated using the ICC, and consensus, estimated from the mean of the absolute value of the difference between raters/methods (9). Scores were compared for the DACTS total scale and for each subscale. Sensitivity and specificity were calculated to assess classification accuracy.

Results

Self-report data were missing for 9 of 16 teams. The maximum number of missing items at a site was two and the mean was .81 (SD=.83). Because phone raters gathered missing data during or immediately after the interview, there was no missing data for the phone assessment. DACTS total and subscale scores were calculated using the mean of non-missing items.

Phone-based fidelity reliability was generally good. Interrater reliability (consistency) was very good for the Total (ICC=.98), Human Resources (ICC=.97) and Nature of Services subscales (ICC=.97), and adequate for the Organizational Boundaries subscale (ICC=.77) (see Table 1). Absolute differences between raters were small, indicating good consensus for the total DACTS (mean difference = .04, differences < .25, 5% of scoring range, for 16/16 sites), Human Resources (mean difference = .05, differences < .25 for 15/16 sites), Organizational Boundaries (mean difference = .06, differences < .25 for 16/16 sites), and Nature of Services subscales (mean difference = .07, differences < .25 for 15/16 sites).

Self-reported fidelity reliability varied by subscale. Interrater reliability (consistency) was acceptable for the Total DACTS (ICC=.77) and Nature of Services subscale (ICC=.86), but below recommended standards for the Organizational Boundaries (ICC=.61) and Human Resources subscales (ICC=.47) (see Table 1). Absolute differences between raters (consensus) were small to medium for the total DACTS (mean difference = .14, differences < .25 for 13/16 sites) and Organizational Boundaries subscale (mean difference = .13, differences < .25 for 13/16 sites), but were somewhat larger for the Nature of Services

(mean difference = .20, differences < .25 for 11/16 sites) and Human Resources subscales (mean difference = .25, differences < .25 for 10/16 sites).

Self-reported fidelity was an accurate/valid predictor of phone-based fidelity (i.e., demonstrated acceptable levels of consistency and consensus) (Table 1). ICCs indicated moderate to strong agreement (consistency) for the Total (ICC=.86) and Nature of Services subscales (ICC=.92), and adequate agreement on the Human Resources (ICC=.74) and Organizational Boundaries subscales (ICC=.71). Absolute differences between self-report and phone (consensus) tended to be small for the Total DACTS (mean difference = .13, differences < .25 for 15/16 sites) and Organizational Boundaries subscale (mean difference = .07, differences < .25 for 15/16 sites) but were somewhat larger for the Nature of Services (mean difference = .20, differences < .25 for 12/16 sites) and Human Resources subscales (mean difference = .15, differences < .25 for 10/16 sites) (see Table 1). Interestingly, self-report DACTS Total scores underestimated phone scores for 12 of 16 sites.

Self-report sensitivity and specificity were calculated to determine classification accuracy when making dichotomous judgments (e.g., ACT vs. non-ACT). Phone DACTS total scores served as the criterion. Teams scoring 4.0 or higher were classified as meeting ACT fidelity standards. Self-report fidelity had a sensitivity of .77, a false positive rate of 0.0, a specificity of 1.00, a false negative rate of .23, and an overall predictive power of .81 in predicting phone fidelity.

Item-level analyses were undertaken to help identify potential problem items. Mean absolute differences between self-report raters (reliability) and between self-report and phone consensus ratings (validity) were examined to identify highly discrepant items. Mean absolute differences exceeding .25 were found between phone raters for 7 items, vocational specialist on team (.46), time unlimited services (−.42), contacts with informal support system (−.38), staff continuity (−.37), dual diagnosis model (−.33), intake rate (−.31), and nurse on team (−.31) and between phone and self-report consensus ratings for 5 items, dual diagnosis model (−.76), vocational specialist on team (−.63), contacts with informal support system (−.44), 24 hour crisis services (−.38), and peer counselor on team (−.37). Most differences were due to either site errors in reporting data or rater errors in judging correctly-reported data. Changes to the protocol could be identified to improve two items—24 hour crisis services and presence of trained vocational specialist on the team (e.g., specify percent of clients calling crisis who speak directly to an ACT team member).

Discussion

The results provide preliminary support for the reliability and validity of the DACTS total scale calculated solely from self-reported data. When restricted to the total DACTS score, which is used to make overall fidelity decisions, rating consistency and consensus were good to very good between self-report raters (i.e., reliability) and between phone and self-report mean rater scores (i.e., validity). In addition, self-reported assessment was accurate, agreeing with the phone assessment total DACTS score within .25 scale points (5% of the scoring range) for 94% of sites, and had a sensitivity of .77, a specificity of 1.0, and overall predictive power of .81 for dichotomous judgments. Moreover, there was no evidence for

inflated self reporting. Self-report fidelity underestimated phone fidelity for most sites. These findings are in contrast to prior research that self-report data is generally less accurate and positively biased (10,11,12), especially when data are subjective or require nuanced clinical ratings. However, prior research has not tested a self-report protocol specifically created to improve accuracy by reducing subjectivity and deconstructing complicated judgments. Moreover, prior studies allowed self-reporters to score themselves. In our study, the self-report data were scored by independent raters.

Results for the DACTS subscales were mixed. Although reliability was very good for the Nature of Services subscale, it ranged from low acceptable to unacceptable for the Organizational Boundaries and Human Resources subscales. A similar pattern was found for validity; excellent for the Nature of Services subscale, but acceptable to low acceptable for the other two subscales. Differences in the number of problem items across subscales appear to underlie the lower reliability and validity (2/10 problem items for Nature of Services vs. 4/11 problem items for Human Resources and 3/7 for Organizational Boundaries). Interestingly, initial results from an ongoing study comparing onsite, phone and self-report fidelity that include the modifications identified earlier for two problem items, show improvements in subscale reliability and validity.

The study had several limitations. The sites were previously certified ACT teams in a single state with clearly defined standards for ACT certification, were mature teams with extensive experience in fidelity assessment, had undergone prior technical assistance and had a history of generally good prior fidelity, and were willing to commit the time required for a detailed self-assessment. These concerns limit both the generalizability and the range of fidelity explored. Also, the carefulness, comprehensiveness, and accuracy of the self-report data may have been impacted, positively (data will be checked) or negatively (data can be fixed later) by the requirement for concurrent phone assessment. In addition, we used phone fidelity as the criterion fidelity measure, based on prior research demonstrating evidence of validity (6). However, future research is needed to confirm that self-report is valid when compared to onsite fidelity. Relatedly, because phone and self-report shared a rating source, conclusions about agreement are limited to comparisons across collection methods and not across collection sources (e.g., independent data collection by onsite rater). Also, the DACTS includes several objective items that do not require clinical judgment, potentially limiting generalizability to fidelity scales with similar types of items.

Despite limitations, the study provides preliminary evidence for the viability of self-report ACT fidelity assessment. However, there are several caveats to its use. First, self-report fidelity is most clearly indicated for gross, dichotomous judgments of adherence using the total scale, and is likely not as useful or sensitive for identifying problems at the subscale or individual item level (e.g., for quality improvement). Second, while there was some time savings for the assessor when using self-report, there was little savings for the site beyond not having to do the phone interview. Moreover, there was a cost in missing data and in lower overall reliability and validity. Third, as is true for phone fidelity, self report cannot, and should not, replace onsite fidelity. Instead, all three methods could be integrated into a hierarchical fidelity assessment approach (6,13). For example, onsite fidelity is likely needed when the purpose of the assessment is quality improvement, and for assessing new

teams and teams experiencing a major transition or trigger event (e.g., high team turnover, decrement in outcomes). Self-report fidelity is likely appropriate to assess overall adherence for stable, existing teams with good prior fidelity. In addition, self-assessment probably is most appropriate as a screening assessment, confirming that prior levels of fidelity remain stable, rather than as the sole indicator of changed performance. That is, evidence for substantial changes at a less rigorous level of assessment (e.g., self-report), will require follow-up assessment using more rigorous methods (e.g., phone and then onsite) to confirm changes.

Acknowledgments

This study was funded by an IP-RISP grant from the National Institute of Mental Health (R24 MH074670; Recovery Oriented Assertive Community Treatment). We appreciate the assistance of others in the collection of data for this study.

References

1. Mancini AD, Moser LL, Whitley R, et al. Assertive community treatment: Facilitators and barriers to implementation in routine mental health settings. *Psychiatric Services*. 2009; 60:189–195. [PubMed: 19176412]
2. Phillips SD, Burns BJ, Edgar ER, et al. Moving assertive community treatment into standard practice. *Psychiatric Services*. 2001; 52:771–779. [PubMed: 11376224]
3. McHugo GJ, Drake RE, Teague GB, et al. The relationship between model fidelity and client outcomes in the New Hampshire Dual Disorders Study. *Psychiatric Services*. 1999; 50:818–824. [PubMed: 10375153]
4. McGrew JH, Bond GR, Dietzen LL, et al. Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology*. 1994; 62:670–680. [PubMed: 7962870]
5. Evidence-based Practice Reporting for Uniform Reporting Service and National Outcome Measures Conference; Bethesda, MD. September 2007;
6. McGrew J, Stull L, Rollins A, et al. A comparison of phone-based and onsite-based fidelity for Assertive Community Treatment (ACT): A pilot study in Indiana. *Psychiatric Services*. 2011; 62:670–674. [PubMed: 21632738]
7. Teague GB, Bond GR, Drake RE. Program fidelity in assertive community treatment: development and use of a measure. *American Journal of Orthopsychiatry*. 1998; 68:216–32. [PubMed: 9589760]
8. McHugo GJ, Drake RE, Whitley R, et al. Fidelity outcomes in the national implementing evidence-based practices project. *Psychiatric Services*. 2007; 58:1279–1284. [PubMed: 17914003]
9. Stemler, SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability; Practical Assessment, Research & Evaluation. 2004. p. 9retrieved June 8, 2010 from <http://PAREonline.net/getvn.asp?v=9&n=4>
10. Adams AS, Soumerai SB, Lomas J, et al. Evidence of self-report bias in assessing adherence to guidelines. *International Journal for Quality in Health Care*. 1999; 11:187–192. [PubMed: 10435838]
11. Lee N, Cameron J. Differences in self and independent ratings on an organizational dual diagnosis capacity measure. *Drug and Alcohol Review*. 2009; 28:682–684. [PubMed: 19930024]
12. Martino S, Balli S, Nich C, et al. Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. *Psychotherapy Research*. 2009; 19:189–193.
13. McGrew, J.; Stull, L. Alternate methods for fidelity assessment. Gary Bond Festschrift Conference; Indianapolis, IN. September 23, 2009;

Table 1

Dartmouth Assertive Community Treatment Scale (DACTS) total scale and subscale self-report and phone-based reliability and validity

Reliability comparisons for phone-based assessment	Rater 1		Rater 2		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
DACTS total scale	4.22	.25	4.20	.28	.04	.00 – 0.11	.98
Organizational Boundaries Subscale	4.58	.14	4.57	.14	.06	.00 – 0.14	.77
Human Resources Subscale	4.27	.35	4.30	.36	.05	.00 – 0.27	.97
Nature of Services Subscale	3.91	.41	3.84	.46	.07	.00 – 0.40	.97

Reliability comparisons for self-report assessment	Rater 3		Rater 4		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
DACTS total scale	4.16	.27	4.11	.26	.14	.00 – 0.41	.77
Organizational Boundaries Subscale	4.49	.20	4.53	.21	.13	.00 – 0.42	.61
Human Resources Subscale	4.27	.39	4.21	.28	.25	.00 – 0.91	.47
Nature of Services Subscale	3.72	.50	3.76	.48	.20	.00 – 0.60	.86

Validity Comparisons	Self-Report consensus		Phone consensus		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
DACTS total scale	4.12	.27	4.21	.27	.13	.00–.43	.86
Organizational Boundaries Subscale	4.53	.15	4.56	.12	.08	.00–.29	.71
Human Resources Subscale	4.22	.31	4.29	.34	.15	.00–.64	.74
Nature of Services Subscale	3.72	.49	3.87	.47	.20	.07–.50	.92

Note: DACTS scores range from 1 to 5, 5=full implementation.