A comparison of phone-based and on-site assessment of fidelity for Assertive Community Treatment in Indiana

John H. McGrew, Ph.D.
Laura G. Stull, M.S.
Angela L. Rollins, Ph.D.
Michelle P. Salyers, Ph.D.
Lia J. Hicks, M.B.A.

All of the authors except Ms. Hicks are affiliated with the Department of Psychology at Indiana University–Purdue University Indianapolis. They are also with the Center on Implementing Evidence-Based Practice, Health Services Research and Development, Roudebush Department of Veterans Affairs Medical Center, Indianapolis. Ms. Hicks is with the Adult and Child Mental Health Center, Indianapolis. Send correspondence to Dr. McGrew at IUPUI, 402 N. Blackford St., LD 124, Indianapolis, IN 46202 (e-mail: jmcgrew@iupui.edu).

Running Head: Phone-based fidelity assessment

Acknowledgements: 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.

Disclosures: None for any author.

Accepted version; Final version published as McGrew, J. H., Stull, L. G., Rollins, A. L., Salyers, M. P., Hicks, L. (2011). A comparison of phone-based and on-site assessment of fidelity for assertive community treatment in Indiana. Psychiatric Services, 62(6), 670-674.

Abstract

Objective: Agencies are being challenged by the burden of monitoring the faithful implementation of evidence based practices. This study investigated the reliability, validity, and role of rater expertise in phone-administered fidelity assessment. Methods: Phone-based and onsite fidelity were compared for 23 ACT teams. A phone protocol for the Dartmouth Assertive Community Treatment Scale was developed. Phone-based raters included an experienced rater plus either a naïve (no prior experience with fidelity) or consultant rater (prior experience with the site). Results: Phone assessment was reliable and valid compared to onsite, according to interrater consistency (intraclass correlation) and consensus (mean rating differences). Phone assessment agreed with onsite within .1 scale point (2% of scoring range) for 83% of sites and within .15 scale point for 91% of sites. Results were unaffected by rater. Conclusions: Phone-based fidelity could potentially address widespread concerns regarding costs of ongoing program monitoring necessary for maintaining high quality services.

A critical concern for mental health services is the problem of deficient implementation of evidence based practices (1, 2) and the corresponding decrement in program outcomes (3, 4). One accepted strategy to improve implementation is to verify program fidelity (5, 6). However, with the increasing number of evidence based practices, the need to conduct fidelity measurement has begun to place a very high burden on agencies charged with ensuring service quality. For example, the current standard fidelity instrument for ACT, the Dartmouth Assertive Community Treatment Scale (DACTS; 7), requires one day for the onsite visit and another day to score and write the report for quality improvement feedback.

In response to these and related problems, a 2007 national task force met to identify alternative approaches for ensuring quality (8). Among the strategies discussed was alternative fidelity methods such as phone-administered assessments. Although phone-administered fidelity has been used successfully in predicting outcomes (9), no research has validated phone-administered assessment compared to onsite assessment. We examined the interrater reliability and concurrent validity of phone-administered fidelity as applied to Assertive Community Treatment. A secondary question asked whether validity and reliability are higher when using raters with prior experience with fidelity assessment or experience with the site.

Methods

Thirty-two ACT teams in Indiana were invited to participate; 23 (71.9%) agreed. All programs had been operating for at least one year, adhered to the Indiana ACT standards, and were receiving annual onsite fidelity assessments from the ACT Center of Indiana (10). The study took place between October 2008 and March 2010.

Onsite Fidelity. The 28-item DACTS (11) assesses fidelity to ACT along three dimensions:

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=full implementation, 1=not implemented). Mean item scores of 4 and above are considered characteristic of established ACT teams. The DACTS has excellent interrater reliability (11) and can differentiate between ACT and other types of intensive case management (7).

Phone-based Fidelity. A phone protocol was developed based on prior phone-based assessment experience (13), and incorporated two key principles. (1) Convert subjective, global questions into molecular, objective data, e.g., replace a global evaluation of responsibility for treatment with a review of specific clients receiving services outside the ACT team. (2) Use of a tabular format, e.g., the staffing table included information about role and hours on team, supervisor, team meeting attendance, turnover, and vacancies, providing data for scoring 11 DACTS items. The DACTS phone protocol included nine tables (plus detailed instructions for completion): staffing, caseload and discharges, client admissions, client hospitalizations, client contact hours and frequency, services received outside of ACT, engagement mechanisms, substance abuse treatment, and miscellaneous (program meeting, practicing team leader, crisis services, and work with informal supports).

Fidelity Assessment Experience. Participants answered the following questions about the phone assessment process: time to complete and prepare and open-ended questions relating to preparation activities, assessment burden/helpfulness, and suggestions for improvement.

Phone interview. To insure that information on burden was not confounded by prior completion of the onsite visit, DACTS phone interviews were conducted before (M=6.78 days), but no more than one month prior to, the onsite visit. However, due to scheduling difficulties, the onsite

interview occurred 49 days after the phone interview at one site and 12 days before the phone interview at another. Sites received a copy of the phone protocol for review two weeks before the phone interview. Team leaders consulted clinical and other program records to complete tables and were encouraged to contact the research team with questions.

Phone interviews were conducted by two raters. To assess impact of rater expertness and prior onsite experience, we systematically varied raters. For half of the interviews, raters were not directly involved with training/consultation of ACT teams: the first author (experienced rater), who has extensive experience conducting both phone and onsite fidelity assessment and a research assistant with no prior experience with fidelity assessment (naïve rater). The remaining interviews were conducted by the first author and the assigned ACT Center consultant, who was familiar with the site. Sites were assigned a second rater (consultant or naive) using quota sampling, stratified by population density (rural vs. urban) and consultant (there were four consultants). Assignment was balanced across the two strata. Adjustments were required when teams declined to participate or had scheduling conflicts. Overall, consultants rated about half of their teams (57%, 33%, 43%, and 50%) and the naïve rater conducted calls with about half of the urban (53%) and rural teams (50%).

To test the validity of phone interviews with minimal staff burden, the team leader was the only site participant. The interview focused on reviewing completed tables for accuracy. In the three cases where the tables were not completed before the call, the interview focused on completing the tables together with the team leader.

Raters independently scored the fidelity items and then discussed their scores to come to consensus. Research team raters based their scores solely on the answers given during the phone interview. Consultants' scores could be informed by knowledge of ACT team operation from prior contact.

Onsite interview. The ACT Center consultant assigned to the team conducted the onsite fidelity assessments. Consultants mailed a checklist of fidelity assessment items/activities to team leaders prior to the site visit (e.g., team roster, interviews with specific staff members). The phone protocol was not used for the onsite interview. The onsite visit typically involved a one hour observation of the daily team meeting, 1.5 – 2 hours to interview the program leader, a half-hour interview with the substance abuse specialist, 1-2 hours shadowing team members in the community and interviewing clients, 2-3 hours of chart and other record reviews, and a half-hour for wrap-up questions with the program leader. Consultants completed DACTS scoring within 5 working days of the site visit and were free to contact program leaders to clarify data if needed.

Consultants received extensive initial training on the DACTS, reviewed the DACTS protocol and scoring at an all-day training workshop annually, and had at least two years experience conducting DACTS assessments. Questions/issues regarding DACTS scoring were addressed through email contact and a bi-weekly meeting.

Although we could not verify onsite reliability, during the first three years of the contract and throughout Indiana's participation in the National Implementing Evidence-based Practices Project, all onsite assessments were completed using two raters, with nearly perfect interrater reliability, intraclass correlation (ICC)=.99 (15).

Analysis. We adopted Stemler's (12) suggestion to use both consensus (raters agree closely and adopt common meaning of scale) and consistency estimates (raters rank sites similarly and are self consistent in their application and understanding of scale) of interrater reliability. Consistency may be high when consensus is low. For phone interviews, interrater consistency was calculated using the ICC. Interrater reliability was calculated across all rater pairs (experienced vs. second rater) and separately for experienced vs. consultant and experienced vs. naïve rater. Interrater

consensus was indexed using the mean and range of the absolute value of the difference between raters. Concurrent validity between phone and onsite ratings was calculated using ICCs (consistency) and the mean and range of the absolute value of the difference between ratings (consensus). Phone and onsite interview scores were compared for DACTS total and subscale scores and when using different raters/ rater pairs (i.e., consensus, naïve rater, consultant rater, phone experienced rater). Calculations for all ICCs followed Shrout and Fleiss (13) Model 2 and used two-way random effects ANOVA with absolute agreement. ICCs above .90 are very good, above .80 are acceptable, and above .70 are adequate for exploratory research (14).

Results

Is phone-based fidelity assessment reliable? Focusing first on the experienced vs. second rater comparisons, the ICCs indicated high levels of reliability (consistency agreement) for Total (ICC=.92), Human Resources (ICC= .93) and Nature of Services (sub)scales (ICC=.91), and adequate reliability for Organizational Boundaries (ICC=.78) (see Table 1). Absolute differences between raters also were small, indicating consensus, for Total, Organizational, and Human Resources subscales using both the mean (.07, .08, .11) and the range (largest discrepancy < .3). Absolute differences for Nature of Services were slightly larger (mean=.18), with the largest discrepancy of .50. There was no discernable impact of prior experience with phone assessment or with the site on consistency (ICCs ranged from .91 to .92) or consensus agreement (mean differences ranged from .06 to .07).

Are phone-based fidelity ratings valid? The onsite and phone-based ratings demonstrated consistency and consensus (Table 1). Focusing first on consensus ratings, ICCs indicated strong agreement between phone and onsite (consistency) for the Total (ICC=.87), Human Resources (ICC= .88) and Nature of Services subscales (ICC=.87), and lower agreement for Organizational

Boundaries (ICC=.69). With one exception, the absolute differences between phone and onsite (consensus) tended to be small for the scales/subscales as measured using the mean (mean differences < .14) and the range (largest discrepancy < .33). However, Nature of Services had a discrepancy between phone and onsite at one site of .5. Phone and onsite ratings for the Total scale differed by .10 or less for 19 sites (83%) and by .15 points or less for 21 sites (91%).

There was a small effect of phone rater on consistency (ICC) but not consensus (Table 1). The ICC between phone and onsite was highest when the consultant did both ratings (.92), was similar to the consensus rating when phone ratings were made by the experienced rater (ICC=.86) and lowest when made by the naïve rater (ICC=.79).

Fidelity phone calls averaged 71.5 +/- 20.5, and ranged from 40 to 111 minutes. Site preparation time for the phone interview, averaged a workday (M=7.6+/-5.9 hours) and ranged from 1.8 hours to 25.0 hours. Preparation time was impacted by availability of electronic medical records and variability in record keeping (e.g., ongoing tracking of clinical activities). Universally, team leaders liked the phone assessment, particularly the table format, felt it was straightforward and rated it either less difficult or comparable to preparing for onsite assessment. However, they expressed concerns that phone assessment should not be the exclusive method of fidelity assessment, worried that it limits contact with consultants, reducing training opportunities and ecological validity of assessment, and suggested including other team members during the assessment, especially the substance abuse specialist.

Discussion

The results indicate that phone assessment is reliable and valid. Phone assessment also appeared to be unbiased, i.e., neither over- or under-estimating onsite scores, and accurate, agreeing with the onsite assessment within .1 scale point (2% of the scoring range) for 83% of sites and

within .15 scale point for 91% of sites. These results provide strong support for the usefulness of phone fidelity assessment.

Surprisingly, rater prior experience, either with phone assessment or the site, had no discernable impact on reliability and only a minor and ambiguous impact on validity. Increased phone-onsite consistency for the consultant rater probably reflected method variance (same rater for both assessments), rather than increased accuracy. Moreover, increased consistency for the experienced rater was not matched by increased consensus with onsite scores. Two factors may explain the small impact of rater: the minor role of the interview(er) in the phone assessment process and the success of the phone protocol in creating an objective, molecular format for gathering fidelity data. For example, protocol preparation time averaged nearly a day, whereas the phone interview took about an hour. Because the phone interview largely focused on verifying the information already tabulated by the team leader, the rater's role during the interview was less of an expert observer, and more of an auditor ensuring self-report accuracy. These results suggest that self-report, if based on clear, objective criteria, may be a useful adjunctive method for fidelity assessment.

The study has several limitations. All the sites were in one state, limiting generalizability, and were certified as ACT teams, limiting the range of fidelity explored. Team leaders were not blind to the study hypotheses, possibly biasing the results, and many were experienced with fidelity reviews, potentially underestimating preparation time required for those naïve to fidelity measurement. Onsite fidelity was conducted using a single rater, providing no opportunity to verify interrater reliability, and the experienced rater was limited to a single individual, limiting generalizability. Finally, we did not measure onsite assessment preparation or site visit time, limiting our ability to compare burden levels.

Despite limitations, the study provides strong evidence for the viability of phone-based ACT fidelity assessment. Further work is needed to examine phone fidelity with other evidence based practices (e.g., supported employment) and with new teams or lower fidelity teams in situations when reimbursement is contingent on high scores. There are several caveats to phone fidelity. For example, although burden to assessors was low, preparation time burden to sites was still high, perhaps prohibitively so. Moreover, phone assessment provides limited opportunity for training and interaction with clients and team members. Thus, phone fidelity cannot and probably should not fully replace onsite fidelity. Instead, both could be integrated into a stepped fidelity assessment approach (15). Onsite fidelity is likely uniquely valuable for assessing teams starting up or experiencing a major transition (e.g., high team turnover). Phone fidelity is likely ideal for stable, mature teams and for frequent check-ins. Future work should explore the relative uses of both methods.

References

1. Phillips SD, Burns BJ, Edgar, ER, et al: Moving assertive community treatment into standard practice. Psychiatric Services 52: 771-779, 2001

2. Drake RE, Essock SM, Shaner A, et al: Implementing dual diagnosis services for clients with severe mental illness. Psychiatric Services 52: 469-476, 2001

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 50: 818-824, 1999

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 62: 670-680, 1994

5. Mancini AD, Moser LL, Whitley R, et al: Assertive community treatment: Facilitators and barriers to implementation in routine mental health settings. Psychiatric Services 60: 189-195, 2009

6. Bond GR, Williams J, Evans L, et al: Psychiatric Rehabilitation Fidelity Toolkit. Cambridge, MA, Human Services Research Institute, 2000

7. Teague GB, Bond GR, Drake RE: Program fidelity in assertive community treatment: development and use of a measure. American Journal of Orthopsychiatry 68: 216-32, 1998

8. Evidence-based Practice Reporting for Uniform Reporting Service and National Outcome Measures Conference, Bethesda, MD, September 2007

9. McGrew JH, Griss ME: Concurrent and predictive validity of two scales to assess the fidelity of implementation of supported employment. Psychiatric Rehabilitation Journal 29: 41-47, 2005

10. Salyers MP, McKasson RM, Bond GR, et al: The role of technical assistance centers in implementing evidence-based practices: lessons learned. American Journal of Psychiatric Rehabilitation 10: 85-101, 2007

11. McHugo GJ, Drake RE, Whitley R, et al: Fidelity outcomes in the national implementing evidence-based practices project. Psychiatric Services 58: 1279-1284, 2007

12. Stemler SE: A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation 9: retrieved June 8, 2010 from http://PAREonline.net/getvn.asp?v=9&n=4, 2004

13. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86: 420-428, 1979

14. Lombard M, Snyder-Duch J, Bracken, CC: Content analysis in mass communication: assessment and reporting of intercoder reliability. Human Communication Research 28: 587-604, 2002

15. McGrew J, Stull L: Alternate methods for fidelity assessment, in Gary Bond Festschrift Conference, Indianapolis, IN, September 23, 2009

Table 1. Dartmouth Assertive Community Treatment Scale (DACTS) total scale and subscale reliability and validity

Reliability comparisons for phone-based assessment	Experienced Rater		Second Phone Rater		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
Total DACTS							
Experienced vs. Second Rater (n=23)	4.30	.17	4.32	.18	.07	.00 – 0.25	.92
Experienced vs. Consultant (n=11)	4.38	.12	4.41	.14	.07	.00 – 0.25	.92
Experienced vs. Naïve (n=12)	4.23	.18	4.23	.17	.06	.00 – 0.14	.91
Organizational Boundaries Subscale[1]	4.73	.15	4.71	.17	.08	.00 – 0.29	.78
Human Resources Subscale[1]	4.34	.23	4.38)	.26	.11	.00 – 0.27	.93
Nature of Services Subscale[1]	3.96	.31	3.98	.30	.18	.00 – 0.50	.91
Validity Comparisons	Consensus Phone		Onsite		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
Total DACTS							
Consensus ratings (n=23)	4.29	.18	4.30	.14	.07	.00 – 0.32	.87
Consultant rater (n=11)	4.41	.14	4.36	.11	.06	.00 – 0.32	.92
Experienced rater (n=23)	4.30	.17	4.30	.14	.07	.00 – 0.25	.86
Naïve rater (n=12)	4.23	.17	4.25	.14	.08	.00 – 0.29	.79
Organizational Boundaries Subscale[2]	4.71	.17	4.72	.17	.09	.00 – 0.29	.69
Human Resources Subscale[2]	4.35	.24	4.34	.26	.11	.00 – 0.27	.88
Nature of Services Subscale[2]	3.93	.29	3.96	.23	.14	.00 – 0.50	.87

Note: DACTS scores range from 1 to 5, 5=full implementation.
Note 1: experienced vs. second rater; Note 2: consensus ratings