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.

Objective: **This study investigated the reliability and validity of a phone-administered fidelity assessment instrument based on the Dartmouth Assertive Community Treatment Scale (DACTS).** ***Methods:*** **An experienced rater paired with a research assistant without fidelity assessment experience or a consultant familiar with the treatment site conducted phone-based assessments of 23 teams providing assertive community treatment in Indiana. Using the DACTS, consultants conducted on-site evaluations of the programs.** ***Results:*** **The pairs of phone raters revealed high levels of consistency [intraclass correlation coefficient (ICC)=.92] and consensus (mean absolute difference of .07). Phone and on-site assessment showed strong agreement (ICC=.87) and consensus (mean absolute difference of .07) and agreed within .1 scale point, or 2% of the scoring range, for 83% of sites and within .15 scale point for 91% of sites. Results were unaffected by the expertise level of the rater.** ***Conclusions:*** **Phone-based assessment could help agencies monitor faithful implementation of evidence-based practices. (*Psychiatric Services* 62: 670–674, 2011)**

Poor implementation of evidence-based practices (1,2) and its toll on program outcomes are a critical concern for mental health services (3,4). One accepted strategy to improve implementation is to verify program fidelity (5,6). However, as the number of evidence-based practices increases, 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 assertive community treatment (ACT), the Dartmouth Assertive Community Treatment Scale (DACTS) (7), requires one day for the on-site visit and another day to score and write the report for quality improvement feedback.

In 2007 a national task force met to identify alternative approaches for ensuring quality (8). Among the strategies discussed was the use of alternative fidelity assessment methods such as phone-administered assessments. Phone-administered fidelity assessment has been used successfully to predict consumer outcomes—for example, employment in a supported-employment program (9)—but no research has validated its use compared with on-site assessment. We examined the interrater reliability of pairs of raters who independently conducted phone assessments of 23 ACT programs. Validity of phone-administered fidelity assessment was examined by comparing results of phone assessment with on-site evaluation. We also examined whether validity and reliability were higher when the raters had prior experience with fidelity assessment or prior knowledge of the participating team.

Methods

Thirty-two ACT teams in Indiana were invited to participate, and 23 (72%) agreed. All programs had been operating for at least one year, adhered to the Indiana ACT standards, and were receiving annual on-site fidelity assessments from the ACT Center of Indiana (10). The study took place between October 2008 and March 2010.

On-site assessments using the 28-item DACTS were conducted by the consultant assigned to the team at each ACT site. The DACTS (11) assesses fidelity to ACT along three dimensions: human resources (for example, hours of psychiatrist time assigned to teams), organizational boundaries (for example, use of explicit admission criteria), and nature of services (for example, use of in vivo

All of the authors except Ms. Hicks are affiliated with the Department of Psychology, Indiana University–Purdue University Indianapolis (IUPUI), and 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).

services). Items are rated using a 5-point scale, from 1, indicating no implementation, to 5, indicating full implementation. 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).

Consultants mailed a checklist of needed items or activities to team leaders before their visit (for example, have team roster ready, plan interviews with specific staff members). The on-site visit, usually about a day long, typically involved one hour spent observing the daily team meeting, one-and-a-half to two hours interviewing the program leader, a half-hour interviewing the substance abuse specialist, one to two hours shadowing team members in the community and interviewing clients, two to three hours reviewing charts and other records, and a half-hour asking the program leader some wrap-up questions. Consultants completed DACTS scoring within five working days of their visit and were free to contact program leaders to clarify data if needed. In all but one case, the on-site assessments were conducted after the phone assessments (see below).

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. They were able to ask questions or discuss issues regarding DACTS scoring through e-mail contact and at biweekly meetings with one another and their supervisor (also a consultant).

Although we could not verify on-site reliability of the DACTS in this study, results from the first three years of the state contract and throughout Indiana's participation in the National Implementing Evidence-based Practices Project found nearly perfect interrater reliability between two raters on all on-site assessments (intraclass correlation coefficient [ICC]=. 99) (11).

A protocol based on the DACTS to conduct fidelity assessment by phone was developed based on prior experience and incorporated two key principles (9). First, subjective, global questions were modified to elicit molecular, objective data. For example, instead of asking the team leader to provide an estimate of the percentage of client psychiatric admissions which involved the team, the phone protocol required the team leader to provide a list of the last 10 psychiatric admissions with a written explanation of team involvement for each one. Second, the instrument relied on tables, nine in all, to report data about most DACTS items, including 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). The staffing table alone included information about role and hours of team members, qualifications of the supervisor, team meeting attendance, turnover, and vacancies, providing information to score 11 items.

Team leaders at the 23 ACT sites were sent a copy of the phone protocol for review two weeks before the phone interview and were asked to complete the tables using any clinical and other program records available. They were encouraged to contact the research team with questions. To minimize staff burden only team leaders participated in completing the protocol.

Phone interviews were conducted by the first author, who has extensive experience conducting both phone and on-site fidelity assessment, and either a research assistant with no prior experience with fidelity assessment (naïve rater) or the consultant assigned to the ACT site. The call was attended by both raters together. Sites were assigned the second rater (consultant or naive) by quota sampling, stratified by population density (rural versus urban) and consultant. Assignment was balanced across the two strata. Rater assignments were adjusted when teams declined to participate or had scheduling conflicts. Overall, consultants rated about half of the teams with which they worked (57%, 33%, 43%, and 50%), and the naïve rater conducted calls with about half of the urban (53%) and rural teams (50%).

The phone interview focused on reviewing completed tables for accuracy. In the three cases in which team leaders had not completed the tables before the raters called, the interview focused on working together to complete them.

Participants were also asked about the time required to complete the tables, answered open-ended questions concerning the burden or helpfulness of the assessment, and made suggestions for improvement.

To ensure that information on burden was not confounded by prior completion of the on-site visit, DACTS phone interviews were conducted before (mean of 6.78 days earlier) but no more than one month before the on-site visit. However, because of scheduling difficulties, the on-site interview occurred 49 days after the phone interview at one site and 12 days before it at another.

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

We adopted a suggestion by Stemler (12) to use both consensus (raters agree closely and adopt common meaning of the scale) and consistency estimates (raters rank sites similarly and are self-consistent in their application and understanding of the 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 versus second rater) and separately for pairs in which the second rater was experienced or naïve. Interrater consensus was indexed by the mean and range of the absolute value of the difference between raters.

Concurrent validity between phone

and on-site ratings was examined for consistency by using ICCs to compare consensus ratings by phone raters with ratings by on-site consultants. Consensus among the phone-assessment raters and the on-site raters was determined by calculating the mean and range of absolute differences in their scores. Scores were compared for DACTS total score and subscale scores for each type of rater pair. Calculations for all ICCs followed model 2 of Shrout and Fleiss (13) and used two-way random-effects analysis of variance with absolute agreement. ICCs above .90 are very good; above .80, acceptable; and above .70, adequate for exploratory research (14).

Results

High levels of reliability (consistency agreement) were found between the experienced and second raters for total DACTS (ICC=.92) and for human resources (ICC= .93) and nature of services (ICC=.91) subscales (Table 1). Reliability was adequate (ICC=.78) for the organizational boundaries subscale. Mean absolute differences between phone raters also were small, indicating consensus, for the total DACTS (.07) and for organizational (.08) and human resources (.11) subscales. The largest discrepancy in the ranges of absolute differences for those scales was less than .3, which is less than 10% of the maximum possible difference. For the nature of services subscale, the mean absolute difference (.18) was slightly larger, and the discrepancy in the range of absolute differences (.50) was the highest of any subscale. Comparisons between the two types of rater pairs on both consistency (experienced versus naïve, ICC=.91, and experienced versus consultant, ICC=.92) and consensus (.06 and .07, respectively) suggest that prior experience with phone assessment or with the treatment site had no discernible impact.

The on-site and phone-based ratings demonstrated consistency and consensus (Table 1). Strong agreement was found between consensus ratings by phone raters and on-site ratings for the total DACTS (ICC= .87) and for the human resources (ICC=.88) and nature of services (ICC=.87) subscales. Lower agreement was found for organizational boundaries (ICC=.69).

Absolute differences between phone and on-site ratings, a measure of consensus, tended to be small for both the total scale and the subscales, as measured by mean absolute differences of .14 or less. Discrepancies in the range of absolute differences were no greater than .32, with the exception of the .50 discrepancy for nature of services. Overall phone and on-site ratings for the total scale dif-

Table 1

Interrater reliability of a phone-assessment version of the Dartmouth Assertive Community Treatment Scale (DACTS) and a comparison of its validity with onsite use of the DACTS among 23 providers of assertive community treatment in Indiana

	Reliability				Validity						
	Phone rater				Consensus phone		On-site consultant		Absolute differences		
	Experienced		Second								
Scale and rater	M	SD	M	SD	M	SD	M	SD	M	Range	ICC[a]
Total DACTS[b]											
Experienced and second rater (N=23)	4.30	.17	4.32	.18					.07	.00–.25	.92
Experienced and consultant (N=11)	4.38	.12	4.41	.14					.07	.00–.25	.92
Experienced and naïve rater (N=12)	4.23	.18	4.23	.17					.06	.00–.14	.91
Subscale[c]											
Organizational boundaries	4.73	.15	4.71	.17					.08	.00–.29	.78
Human resources	4.34	.23	4.38	.26					.11	.00–.27	.93
Nature of services	3.96	.31	3.98	.3					.18	.00–.50	.91
Total DACTS[b]											
Experienced and second rater (consensus) (N=23)					4.29	.18	4.30	.14	.07	.00–.32	.87
Consultant rater (N=11)					4.41	.14	4.36	.11	.06	.00–.32	.92
Experienced rater (N=23)					4.30	.17	4.30	.14	.07	.00–.25	.86
Naïve rater (N=12)					4.23	.17	4.25	.14	.08	.00–.29	.79
Subscale[d]											
Organizational boundaries					4.71	.17	4.72	.17	.09	.00–.29	.69
Human resources					4.35	.24	4.34	.26	.11	.00–.27	.88
Nature of services					3.93	.29	3.96	.23	.14	.00–.50	.87

[a] Intraclass correlation coefficient
[b] DACTS scores range from 1, not implemented, to 5, fully implemented.
[c] Results are for experienced and second raters (N=23).
[d] Results represent consensus of experienced and second raters (N=23).

fered by no more than .10 points for 19 sites (83%) and by no more than .15 points for 21 sites (91%).

There was a small effect of phone rater on consistency but not on consensus (Table 1). The scores for the on-site assessment were compared to the phone assessment scores for the phone consensus and for each rater separately. Scoring consistency with the on-site assessment (which was completed by the consultant) was highest when compared with the consultant's phone ratings (ICC=.92), was similar to the phone consensus rating when compared to phone ratings made by the experienced rater (ICC=.86), and was lowest when compared to the phone ratings made by the naïve rater (ICC=.79).

Fidelity phone calls ranged from 40 to 111 minutes (mean±SD=71.5± 20.5 minutes). Time spent on preparing for the phone interview ranged from 1.8 to 25 hours (mean±SD of 7.6±5.9 hours). Preparation time was affected by availability of electronic medical records and variability in record keeping, for example, in 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 than or comparable in difficulty to preparing for on-site 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, especially the substance abuse specialist, during the assessment.

Discussion and conclusions

The results indicate that phone assessment of ACT fidelity is reliable and valid. Phone assessment also appeared to be unbiased (neither overestimating nor underestimating on-site scores) and accurate, agreeing with the on-site 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, prior experience of the rater, either with phone assessment or with the site, had no discernible impact on reliability and only a minor and ambiguous impact on validity. Better consistency between ratings for the phone and on-site assessment when a consultant completed both ratings probably reflected method variance, in other words, the same rater did both assessments, rather than increased accuracy when consultants were involved in both evaluations. Moreover, even though pairs without a naïve rater had better consistency, they did not demonstrate increased consensus.

Two factors that may explain the small impact of rater are the minor role of the interview or the interviewer 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 itself 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 that of an expert observer and more that of an auditor ensuring accurate self-reporting. 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, which limits the extent to which the results can be generalized, and were certified as ACT teams, which limits the range of deviations from fidelity that could be explored. Team leaders were not blind to the study hypotheses, possibly biasing the results, and because many were experienced with fidelity reviews, they may have underestimated the amount of time needed by less experienced individuals to prepare for fidelity assessment. On-site fidelity was conducted by a single rater, a design that provided no opportunity to verify interrater reliability. In addition, a single individual served as the experienced rater, which may limit the generalizability of the results. Finally, because we did not measure the time taken to prepare for the on-site assessment or the length of the visit, our ability to compare burden levels of the two assessment protocols is limited.

Despite its limitations, the study provides strong evidence for the viability of phone-based assessment of ACT fidelity. Further work is needed to examine phone-based assessment of fidelity in other evidence-based practices, such as supported employment. Can it be relied upon when the treatment team is new or has a history of having difficulty in maintaining program fidelity, especially if reimbursement is contingent on high scores?

In fact, there are several caveats to phone fidelity. For example, although burden to assessors was low, the burden of preparation at the sites was high, perhaps prohibitively so. Moreover, phone assessment provides limited opportunity for training and interaction with clients and team members. Thus it cannot and probably should not fully replace on-site assessment of fidelity. Instead, both could be integrated into a stepped fidelity assessment approach (15). On-site assessment of fidelity is likely uniquely valuable for assessing teams starting up or experiencing a major transition, such as high turnover. Phone-based assessment of fidelity is likely ideal for stable, mature teams and for frequent check-ins. Future work should explore the relative uses of both methods.

Acknowledgments and disclosures

This study was funded by IP-RISP grant R24 MH074670 from the National Institute of Mental Health and by a contract with the Indiana Family and Social Services Administration, Division of Mental Health and Addiction. The authors appreciate the assistance of the ACT team leaders, as well as Hea-Won Kim, Dave McClow, and Jennifer Wright-Berryman, who helped in the collection of data for this study.

The authors report no competing interests.

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, Mass, 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: 216–232, 1998
8. Evidence-based Practice Reporting for Uniform Reporting Service and National Outcome Measures Conference. Bethesda, Md, September 13, 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 and Evaluation 9, 2004. Available at PAREonline.net/getvn.asp?v=9&n=4
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: Alternative methods for fidelity assessment; in Festschrift for Gary Bond. Indianapolis, Indiana University–Purdue University Indianapolis, Sept 24–25, 2009. Available at www.psych.iupui.edu/users/kjohnson/conweb/festschrift_talks/linkstotalks.html

Submissions for Datapoints Column Invited

Submissions to the journal's Datapoints column are invited. Datapoints encourages the rapid dissemination of relevant and timely findings related to clinical and policy issues in psychiatry. National data are preferred. Areas of interest include diagnosis and practice patterns, treatment modalities, treatment sites, patient characteristics, and payment sources. The analyses should be straightforward, so that the figure or figures tell the story. The text should follow the standard research format to include a brief introduction, description of the methods and data set, description of the results, and comments on the implications or meanings of the findings.

Datapoints columns, which have a one-page format, are typically 350 to 400 words of text with one or two figures. Because of space constraints, submissions with multiple authors are discouraged; submissions with more than four authors should include justification for additional authors.

Inquiries or submissions should be directed to column editors Amy M. Kilbourne, Ph.D., M.P.H. (amy.kilbourne@va.gov), or Tami L. Mark, Ph.D. (tami.mark@thomsonreuters.com).