Adm Policy Ment Health (2016) 43:157–167
DOI 10.1007/s10488-015-0641-1

ORIGINAL ARTICLE

Comparison of Assertive Community Treatment Fidelity Assessment Methods: Reliability and Validity

Angela L. Rollins · John H. McGrew · Marina Kukla · Alan B. McGuire · Mindy E. Flanagan · Marcia G. Hunt · Doug L. Leslie · Linda A. Collins · Jennifer L. Wright-Berryman · Lia J. Hicks · Michelle P. Salyers

Published online: 27 February 2015

Abstract Assertive community treatment is known for improving consumer outcomes, but is difficult to implement. On-site fidelity measurement can help ensure model adherence, but is costly in large systems. This study compared reliability and validity of three methods of fidelity assessment (on-site, phone-administered, and expert-scored self-report) using a stratified random sample of 32 mental health intensive case management teams from the Department of Veterans Affairs. Overall, phone, and to a lesser extent, expert-scored self-report fidelity assessments compared favorably to on-site methods in inter-rater reliability and concurrent validity. If used appropriately, these alternative protocols hold promise in monitoring large-scale program fidelity with limited resources.

Keywords Fidelity · Quality measurement · Implementation · Assertive community treatment

Introduction

Assertive community treatment (ACT) is an effective model of community-based treatment for people with severe mental illnesses (Stein and Test 1980). ACT has been the subject of over 30 randomized controlled trials. Typical outcomes include reduced hospital use, increased housing stability, increased client retention, improvements in level of functioning and quality of life, and increased satisfaction with treatment (Bond et al. 2001, 1995;

A. L. Rollins (✉) · M. Kukla · A. B. McGuire · M. E. Flanagan · L. A. Collins
HSR&D Center for Health Information and Communication, Richard L. Roudebush VAMC, 1481 W. 10th Street, 11-H, Indianapolis, IN 46202, USA
e-mail: alrollin@iupui.edu

M. Kukla
e-mail: mkukla@iupui.edu

A. B. McGuire
e-mail: abmcguir@iupui.edu

M. E. Flanagan
e-mail: meflanag@iupui.edu

L. A. Collins
e-mail: linda.collins4@va.gov

A. L. Rollins · J. H. McGrew · M. Kukla · A. B. McGuire · M. P. Salyers
Department of Psychology, Indiana University Purdue University Indianapolis, Indianapolis, USA
e-mail: jmcgrew@iupui.edu

M. P. Salyers
e-mail: mpsalyer@iupui.edu

M. G. Hunt
Office of Mental Health Operations, VA Northeast Program Evaluation Center, West Haven, USA
e-mail: Marcia.hunt2@va.gov

D. L. Leslie
College of Medicine, Penn State University, University Park, USA
e-mail: dleslie@phs.psu.edu

J. L. Wright-Berryman
School of Social Work, University of Cincinnati, Cincinnati, USA
e-mail: jennifer.wright-berryman@uc.edu

L. J. Hicks
Adult & Child Center, Indianapolis, USA
e-mail: lhicks@adultandchild.org

Herdelin and Scott 1999; Mueser et al. 1998; Phillips et al. 2001; Ziguras and Stuart 2000). In part because of the strong empirical research base, ACT has gained broad acceptance and has been widely disseminated, both in the United States and elsewhere. ACT has been identified as one of six evidence-based practices for the public mental health sector (Drake et al. 2001), has been endorsed in U.S. Governmental reports (President's New Freedom Commission on Mental Health 2003) and by the U.S. Medicaid agency (Clark 2004), and has received vigorous advocacy by the National Alliance on Mental Illness (Allness and Knoedler 1998, 2003; Torrey et al. 2003). In addition, ACT is remarkable for the degree to which its structural and functional features have been articulated (McGrew and Bond 1995), as well as for having a widely-used fidelity scale to assess a team's adherence to an ideal ACT model for staffing and services (Teague et al. 1998).

Although ACT programs are effective, they tend to be difficult to implement accurately. Studies have found wide variability in the degree to which the self-designated "ACT" programs adhere to the original design (McGrew and Bond 1997; McGrew et al. 1994). Unfortunately, variability in implementation, as measured by departures from fidelity (i.e., degree of adherence to an intervention model), can critically affect outcomes in psychosocial programs (Drake et al. 2009; Latimer 1999; McGrew et al. 1994; McHugo et al. 1999) and the sustainability of programs over time (Bond et al. 2014). Given the fact that variability in program implementation is the norm and that uncontrolled variability usually leads to poorer outcomes, there is now broad consensus on the need to verify program fidelity for ACT and for other evidence-based practices. The current standard for assessing ACT fidelity is the dartmouth assertive community treatment scale (DACTS) (Teague et al. 1998), which is administered by experts in the model through on-site visits to the targeted program. The DACTS has good inter-rater reliability and can differentiate between intensive case management models (Teague et al. 1998). A precursor of the DACTS showed a robust correlation between ACT fidelity and reduced hospital use (McGrew et al. 1994), which justifies efforts on the part of health care systems to assess fidelity.

Although there is general agreement on the need for fidelity assessment, there is disagreement about the reliability, validity and cost effectiveness (e.g., assessment burden) of different assessment methods (Bond 2013; McGrew et al. 2013a, b). Due to state and federal fiscal restraints, gold standard on-site visits may not be feasible within large health systems due to their cost in personnel, time, and lost productivity for clinicians. This dilemma has led some to propose alternative methods, such as phone-based and more innovative self-report methods where data are provided by the team but scoring is still done by an expert rater ("expert-scored self-report"). Two studies examined these alternate methods for ACT. In the first study, McGrew et al. (2011) showed that phone-administered DACTS fidelity could be rated reliably and had good agreement with on-site assessment as measured by between-method consistency (i.e., inter-rater reliability) and consensus (i.e., low mean absolute differences in scores). In a second study, they demonstrated that expert-scored self-reported fidelity could be rated reliably and was comparable to phone-administered fidelity, again, as indicated by good consistency and consensus (McGrew et al. 2013b). However, both studies were small, used convenience sampling, took place in a single state and were limited to well-established, stable teams with both good overall fidelity and extensive prior experience in fidelity assessments. These limitations in external validity are problematic because fidelity monitoring, arguably, is equally if not more critical for developing or relatively new teams and for teams that do not have a history of good fidelity. Before such a substantial and economically appealing change to current fidelity assessment methods can be recommended, remote fidelity methods research requires a strong replication in another health system, using a more rigorous, prospective design and randomized sampling of teams with a wider range of ACT fidelity. The current study addresses these needs in applying the alternative fidelity methods in a larger, nationwide study using a rigorous design and sampling framework where uniformly high fidelity to ACT was not expected. Moreover, conclusions about the validity of expert-scored self-report fidelity are problematic because the comparison was to phone-administered fidelity, not on-site fidelity, which is still considered the gold standard. The current study will allow us to directly test the comparability of expert-scored self-report with on-site methods using a prospective study design and blinded raters in each condition.

The primary aims of the current study were to examine the reliability and concurrent validity of three different methods of fidelity assessment through a comparison of expert-scored self-report, phone, and on-site fidelity assessment. Based on high DACTS inter-rater reliability for on-site assessment (ICC = 0.99) (McHugo et al. 1999) and previous work with phone methods, (McGrew et al. 2011; McGrew, Et al. 2013), we expected inter-rater reliability (consistency) above 0.9 and inter-rater consensus as indicated by mean absolute differences of <0.1 (2.5 % of the scoring range). Based on promising results from the prior published work (McGrew et al. 2011), we also expected both remote methods to have excellent consistency with on-site results as indicated by ICCs of at least 0.80 and good consensus with on-site results as indicated by mean absolute differences of <0.2 (5 % of the scoring range). Item-level results were also explored.

Methods

To address these aims, we conducted a cross-sectional, multisite study to compare three different assessment approaches (on-site, phone, and expert-scored self-report fidelity).

Setting

This study took place in the Department of Veteran Affairs (VA) Mental Health Intensive Case Management (MHICM) programs. VA has endorsed ACT as a treatment model of choice and began implementing MHICM programs through a demonstration program in 1987 (Neale et al. 2007). MHICM teams provide intensive community based treatment for veterans, currently growing to 114 MHICM teams at over one hundred healthcare systems across the country. The VA's MHICM teams were used in a large scale study demonstrating ACT's cost-effectiveness (Rosenheck et al. 1995; Rosenheck and Neale 1998) and were included in earlier research to establish the DACTS fidelity scale (Teague et al. 1998).

Sample and Recruitment

We recruited 32 MHICM teams to participate in the study from 2011 to 2013. The pool of potential sites was limited to MHICM teams in existence for 1 year or more ($n = 111$), based on findings from the National Implementing Evidence-based Practices Project that ACT teams typically attain stable fidelity scores by the end of their first year of implementation (McHugo et al. 2007). Teams were selected from the pool based on a stratified random sampling approach with replacement. Of the eligible MHICM teams, 71 % were located within general medicine and surgery (GM&S) facilities, and 29 % were located in neuropsychiatric (NP) settings that historically provided long-term psychiatric inpatient services. To account for possible differences between GM&S and NP sites in the type of Veteran served (e.g., acuity or functioning), we first stratified based on location type. To ensure variability of fidelity scores in the sample and to prevent "spectrum bias"—that a test may have differential predictive validity if tested in an extreme group (Ransohoff and Feinstein 1978)—we also stratified based on team prior year self-reported fidelity scores using overall sample median split. This resulted in four strata: high and low fidelity, GM&S and high and low fidelity NP sites. Teams were selected randomly within each of the four strata. Ten high and 10 low GM&S sites and six high and six low NP sites were invited to participate. When a site declined to participate, we replaced it with the next site on our list.

As part of recruitment efforts, we distributed brochures and made presentations to national MHICM teleconferences to inform MHICM team leaders regarding the study. All procedures were reviewed and approved by the VA's Central IRB and Richard L. Roudebush VA Medical Center's Research and Development Committee.

Measures

Dartmouth Assertive Community Treatment Scale (DACTS)

The DACTS (Teague et al. 1998) is a 28-item scale that assesses degree of fidelity to the ACT model along three dimensions: Human resources (e.g., small caseload, psychiatrist on staff), Organizational boundaries (e.g., explicit admission criteria), and Nature of services (e.g., in vivo services). Each item is rated on a 5-point behaviorally anchored scale, ranging from $1 =$ not implemented to $5 =$ fully implemented. Anchors are item-specific. The DACTS has been shown to discriminate between four types of services (Teague et al. 1998) and is sensitive to change over time in implementation efforts (McHugo et al. 2007). Inter-rater reliability of the on-site DACTS (between two trained raters making a conjoint visit) was found to be 0.99 in the National Implementing Evidence-based Practices Project (McHugo et al. 2007). All three forms of fidelity assessment were based on the DACTS, but used different methods for data collection and rating as described below.

Procedures

We counterbalanced the order of phone and on-site fidelity assessments, such that half the sites were randomly assigned to receive the phone assessment first, and the other half to receive the on-site assessment first. However, because three sites originally assigned to receive phone first rescheduled so that on-site came first, 19 sites received on-site assessments first and 13 sites received the phone assessment first. After agreeing to participate, each MHICM program leader and his/her supervisor were contacted via email and/or phone to begin preparations for their first assessment. Team leaders prepared a set of 10 tables describing objective team composition and activities (e.g., admission criteria, number of veterans receiving services offered, description of recent circumstances around hospitalizations) in advance of the assessment. These tables were used for all three assessments and were expanded from previously established fidelity protocols and from those used in prior studies by McGrew et al. (2011). Revisions included more comprehensive instructions for completing table items (e.g., more precise definitions of graduation or dropout from ACT, rephrasing questions on team meetings that capture the full array of

potential content and frequency that fall below higher DACTS ratings), adding table items to more fully assess team admission criteria and procedures and assertive engagement examples, and adding VA-specific terminology to avoid confusion for the respondent. Expert-scored self-report was completed solely from these tables. Phone and On-site assessments used the tables as the basis for the interviews with the team leader and others. Specific procedures for each method are outlined below. Raters for each assessment type were blinded to the results of other assessment types for that site.

On-Site Fidelity Assessment

Prior to the visit, the MHICM team leader received a checklist of items and data collection sources needed for the on-site visit (e.g., team roster, chart reviews, interviews with specific staff members). On-site visit days were scheduled so that the assessor, the MHICM program leader, and as many key MHICM staff members as possible were present for the visit. Each day-long on-site visit included observation of the daily team meeting, interviews with the program leader, vocational, peer, and substance abuse specialists (if assigned to the team), shadowing team members in the community, and reviewing a random sample of charts and other records. On-site DACTS fidelity assessments used a single rater because inter-rater reliability has been determined to be strong (Salyers et al. 2007). However, to ensure quality control, the first author attended four of the 32 on-site assessments to monitor inter-rater reliability of the on-site approach. The paired ratings for these sites averaged a difference of 0.03 on the total DACTS score and an ICC of 0.99, indicating very high inter-rater consensus and consistency, consistent with previous findings (McHugo et al. 2007).

Phone Fidelity Assessment

Similar to the on-site visits, team leaders completed the fidelity tables in advance. As part of this effort, sites were asked to report de-identified service use data from electronic health records for a random sample of charts. A phone assessor was available to address questions about preparing the fidelity tables or accessing health records before the interview (e.g., suggestions about where to find data for the tables, clarify terminology, etc.). Team leaders were encouraged to fax or email completed tables to the program manager in advance of the phone fidelity call to streamline the phone interview process. To test the validity of phone interviews when conducted with the least possible burden to program staff, we only required the team leader's participation. In addition, team leaders should provide more accurate information than other team members (Bond et al. 2000). Two raters conducted the phone assessments via conference call, independently scored the items on the fidelity scale, and later came to consensus on discrepant items.

Expert-Scored Self-Report Fidelity Assessment

This assessment method used the tables prepared by teams in advance of on-site and phone fidelity as data to score each DACTS item. Two fidelity assessors used the prepared tables without any contact with or clarifications from respondents, independently scored the items, and later came to consensus on discrepant items.

Fidelity Assessors

On-site assessments were conducted by one of two raters who alternated between performing the on-site or phone-based assessment. Three additional raters rotated roles as the second phone-based fidelity rater and one of the two experts scoring self-reported data. All raters were experienced fidelity assessors and participated in a day-long workshop to review DACTS fidelity assessment materials adapted to the VA context. The raters also attended monthly calls throughout the project to discuss scoring rules and protocols for ongoing quality assurance. Fidelity data were entered into a relational database that included embedded computation and scoring capabilities to reduce mathematical errors (e.g., calculation of caseload ratio).

Data Analysis

Descriptive statistics were computed for on-site fidelity data to give an overall sense of how close the MHICM sample adhered to ACT standards. Three indicators were used to assess inter-rater agreement (reliability) and inter-method agreement (concurrent validity): consistency, calculated with the two-way mixed intraclass correlation coefficient (ICC); consensus, estimated from the mean of the absolute value of the difference between raters or methods; and percent agreement (Stemler 2004). We examined inter-rater reliability, i.e., consistency and consensus, for the total score and for each subscale of the DACTS for both phone and expert-scored self-report methods. Concurrent validity, i.e., consistency and consensus, was calculated between all three measures of fidelity for the corresponding DACTS total and subscale scores. Mean differences between the three assessments also were tested using a mixed model (repeated measures) approach, followed by Tukey's pairwise comparisons to adjust for multiple comparisons.

Table 1 Descriptive statistics: DACTS item, subscale, and total scores (n = 32)

		On-site		Phone		Self-report	
		Mean	SD	Mean	SD	Mean	SD
H1	Small caseload	4.41	0.56	4.47	0.57	4.44	0.56
H2	Team approach	3.06	1.22	3.16	1.30	3.28	1.33
H3	Program meeting	3.19	1.31	2.88	1.36	3.00	1.32
H4	Practicing team leader	4.00	1.27	4.13	1.26	3.91	1.28
H5	Continuity of staffing	4.22	0.94	4.13	0.98	4.19	1.00
H6	Staff capacity	4.38	0.75	4.38	0.71	4.53	0.62
H7	Psychiatrist on staff	3.00	1.30	2.88	1.36	2.91	1.40
H8	Nurse on staff	4.78	0.55	4.84	0.45	4.75	0.67
H9	Substance abuse specialist on staff	1.22	0.87	1.31	0.93	1.28	0.92
H10	Vocational specialist on staff	1.22	0.66	1.25	0.76	1.22	0.49
H11	Program size	3.44	1.13	3.44	1.11	3.38	1.10
Human resources mean		3.36	0.40	3.35	0.43	3.35	0.48
O1	Explicit admission criteria	4.13	0.66	3.94	0.72	3.66	0.79
O2	Intake rate	4.97	0.18	4.97	0.18	4.97	0.18
O3	Full responsibility for treatment services	2.78	0.91	2.31	0.82	2.72	1.05
O4	Responsibility for crisis services	2.00	1.63	1.84	1.35	1.88	1.41
O5	Responsibility for hospital admissions	3.31	0.97	3.25	0.80	3.25	0.88
O6	Responsibility for hospital discharge planning	4.53	0.72	4.53	0.76	4.31	0.93
O7	Time-unlimited services	4.63	0.55	4.66	0.55	4.59	0.61
Organizational boundaries mean		3.76	0.38	3.64	0.35	3.62	0.40
S1	In-vivo services	4.75	0.67	4.75	0.62	4.66	0.75
S2	No drop out policy	4.56	0.50	4.53	0.51	4.44	0.50
S3	Assertive engagement mechanisms	3.66	0.70	3.34	0.83	3.00	0.88
S4	Intensity of service	2.53	0.72	2.75	0.76	2.81	0.82
S5	Frequency of contact	1.94	0.56	2.09	0.39	2.19	0.64
S6	Work with support system	2.56	1.27	2.41	1.21	2.63	1.34
S7	Individualized substance abuse treatment	1.63	1.29	1.63	0.98	1.81	1.15
S8	Dual disorder treatment groups	1.13	0.42	1.09	0.53	1.41	1.04
S9	Dual disorders model	2.16	0.57	1.75	0.62	1.94	0.84
S10	Role of consumers on treatment team	1.75	1.59	1.63	1.36	1.72	1.33
Nature of services mean		2.67	0.33	2.60	0.31	2.66	0.40
Total DACTS mean		3.21	0.27	3.15	0.28	3.17	0.31

Results

As seen in Table 1, teams showed modest fidelity to the ACT model as assessed by the on-site method. DACTS score means were below 4.0 (fully implemented) for all subscales and the total scale score. No team in our sample scored a 4.0 or higher on the Total DACTS mean using any fidelity assessment method.

Inter-Rater Reliability

At the DACTS subscale and total scale level, analyses indicated good inter-rater reliability for phone and expert-scored self-report methods using both consistency and consensus measures. As seen in Table 2, consistency, as measured using inter-rater reliability (intraclass correlations) was very good (ICCs = 0.96) for total DACTS scores for both phone and expert-scored self-report and also was high for subscales for each alternative method, with the lowest ICC for the services subscale for the phone method (0.81). Likewise, consensus was high for subscales and total DACTS mean scores with all mean absolute differences falling below 0.20. The mean absolute difference between raters on the total DACTS score was 0.09 and 0.10 for phone and self-report assessment methods, respectively, indicating high consensus between raters for both remote methods.

At the item level, 23 items (82 %) rated via phone and 22 items (79 %) rated via expert-scored self-report had ICCs of 0.80 or above for inter-rater consistency. For

Table 2 Inter-rater reliability indicators for phone and expert-scored self-report methods: intra-class correlations (ICC, average measures), absolute mean differences, and percent agreement between raters by method

		Phone			Self-report		
		ICC	Mean diff	% agree	ICC	Mean diff	% agree
H1	Small caseload	0.98	0.03	96.9	0.71	0.25	81.3
H2	Team approach	1.0	0	100.0	1.0	0.03	96.9
H3	Program meeting	0.97	0.22	78.1	0.83	0.43	71.9
H4	Practicing team leader	0.96	0.22	81.3	0.91	0.47	59.4
H5	Continuity of staffing	0.87	0.34	75.0	0.93	0.25	75.0
H6	Staff capacity	0.88	0.22	81.3	0.82	0.25	75.0
H7	Psychiatrist on staff	0.97	0.16	87.5	0.90	0.28	87.5
H8	Nurse on Staff	0.96	0.03	96.9	0.91	0.09	93.8
H9	Substance abuse specialist on staff	0.95	0.09	93.8	1.0	0	100
H10	Vocational specialist on staff	0.97	0.06	93.8	0.74	0.16	87.5
H11	Program size	0.99	0.06	93.8	0.94	0.22	81.3
Human resources mean		0.96	0.09		0.92	0.13	
O1	Explicit admission criteria	0.38	0.56	46.9	0.60	0.50	56.3
O2	Intake rate	1.0	0	100	0	0	93.8
O3	Full responsibility for treatment services	0.73	0.56	46.9	0.77	0.53	56.3
O4	Responsibility for crisis services	0.90	0.41	75.0	0.95	0.38	62.5
O5	Responsibility for hospital admissions	0.87	0.25	78.1	0.88	0.34	65.6
O6	Responsibility for hospital discharge planning	0.89	0.19	84.4	0.92	0.19	84.4
O7	Time-unlimited services	0.97	0.03	96.9	0.90	0.13	87.5
Organizational boundaries mean		0.81	0.18		0.87	0.16	
S1	In-vivo services	1.0	0	100	1.0	0	100
S2	No drop out policy	0.94	0.06	93.8	0.77	0.19	81.3
S3	Assertive engagement mechanisms	0.52	0.63	53.1	0.49	0.75	43.8
S4	Intensity of service	0.93	0.09	93.8	0.94	0.09	93.8
S5	Frequency of contact	0.88	0.06	93.8	0.96	0.06	93.8
S6	Work with support system	0.96	0.16	87.5	0.98	0.13	84.4
S7	Individualized substance abuse treatment	0.86	0.38	68.8	0.86	0.47	62.5
S8	Dual disorder treatment groups	0.70	0.19	90.6	0.98	0.06	93.8
S9	Dual disorders model	0.60	0.50	53.1	0.81	0.56	46.9
S10	Role of consumers on treatment team	0.97	0.19	84.4	0.88	0.28	84.4
Nature of services mean		0.88	0.16		0.91	0.19	
Total DACTS mean		0.96	0.09		0.96	0.10	

consensus, 17 of the 28 (61 %) items using the phone method and 12 of 28 items (43 %) using expert-scored self-report had mean absolute differences <0.2, indicating high levels of consensus between raters on those items. Percent agreement between raters was 75 % or higher for 23 (82 %) items using the phone method and 19 (68 %) items using expert-scored self-report.

Concurrent Validity

Inter-method agreement between phone, expert-scored self-report, and on-site methods was high for total DACTS score and most subscales using consistency measures. As seen in Table 3, intraclass correlations between phone and on-site methods were 0.96, 0.85, 0.84, and 0.91 for the Human Resources, Organizational Boundaries, Services subscale, and total DACTS score, respectively (all above the 0.80 criterion for phone). Intraclass correlations indicating agreement between self-report and on-site methods were slightly lower: 0.92, 0.66, 0.79, and 0.84 for the Human Resources, Organizational Boundaries, Services subscale, and total DACTS score, respectively. The Organizational boundaries and Services subscales for expert-scored self-report were the only subscales that did not

Table 3 Intermethod agreement using mean absolute differences, range of absolute differences, intra-class correlations (ICC), and percent agreement

		Phone–self-report				Phone–on-site				Self-report–on-site			
		Mean diff	Range	ICC	% agree	Mean diff	Range	ICC	% agree	Mean diff	Range	ICC	% agree
H1	Small caseload	0.16	0, 1	0.86	84	0.13	0, 1	0.89	88	0.16	0, 1	0.86	84
H2	Team approach	0.13	0, 2	0.96	94	0.53	0, 2	0.88	53	0.53	0, 2	0.89	53
H3	Program meeting	0.50	0, 4	0.81	69	0.63	0, 2	0.85	53	0.56	0, 4	0.79	66
H4	Practicing team leader	0.28	0, 2	0.93	78	0.31	0, 2	0.94	72	0.34	0, 2	0.91	75
H5	Continuity of staffing	0.31	0, 2	0.89	72	0.16	0, 1	0.96	84	0.34	0, 2	0.88	69
H6	Staff capacity	0.34	0, 2	0.65	72	0.19	0, 1	0.90	81	0.34	0, 3	0.55	75
H7	Psychiatrist on staff	0.28	0, 3	0.92	84	0.19	0, 2	0.95	88	0.16	0, 3	0.95	91
H8	Nurse on Staff	0.09	0, 3	0.73	97	0.13	0, 2	0.77	91	0.22	0, 3	0.54	88
H9	Substance abuse specialist on staff	0.09	0, 2	0.95	94	0.09	0, 2	0.95	94	0.06	0, 2	0.96	97
H10	Vocational specialist on staff	0.22	0, 3	0.59	88	0.28	0, 4	0.35	88	0.19	0, 3	0.61	88
H11	Program size	0.13	0, 1	0.97	88	0.13	0, 1	0.97	88	0.25	0, 1	0.95	75
Human resources mean		0.14	0, 0.5	0.95		0.14	0, 0.4	0.96		0.17	0, 0.5	0.92	
O1	Explicit admission criteria	0.59	0, 2	0.57	47	0.50	0, 2	0.52	56	0.53	0, 3	0.55	56
O2	Intake rate	0.06	0, 1	−0.07	91	0.06	0, 1	−0.07	94	0.06	0, 1	−0.07	91
O3	Full responsibility for treatment services	0.66	0, 2	0.66	50	0.66	0, 2	0.67	44	0.75	0, 3	0.47	47
O4	Responsibility for crisis services	0.41	0, 2	0.92	66	0.47	0, 2	0.90	69	0.44	0, 4	0.89	72
O5	Responsibility for hospital admissions	0.44	0, 2	0.75	63	0.31	0, 1	0.89	69	0.31	0, 2	0.88	72
O6	Responsibility for hospital discharge planning	0.47	0, 3	0.64	66	0.25	0, 2	0.83	78	0.28	0, 2	0.83	78
O7	Time-unlimited services	0.06	0, 2	0.90	97	0.09	0, 1	0.92	91	0.09	0, 1	0.93	91
Organizational boundaries mean		0.20	0, 0.7	0.86		0.18	0, 0.9	0.85		0.26	0, 1.3	0.66	
S1	In-vivo services	0.09	0, 1	0.95	91	0.31	0, 2	0.56	78	0.34	0, 3	0.51	78
S2	No drop out policy	0.09	0, 1	0.90	91	0.16	0, 1	0.82	84	0.19	0, 1	0.78	81
S3	Assertive engagement mechanisms	0.59	0, 2	0.63	53	0.63	0, 2	0.46	50	0.84	0, 2	0.32	41
S4	Intensity of service	0.19	0, 3	0.82	88	0.47	0, 2	0.64	59	0.53	0, 3	0.43	63
S5	Frequency of contact	0.09	0, 3	0.67	97	0.28	0, 1	0.59	72	0.38	0, 3	0.44	69
S6	Work with support system	0.47	0, 2	0.88	66	0.66	0, 3	0.80	50	0.44	0, 2	0.90	66
S7	Individualized substance abuse treatment	0.56	0, 3	0.74	59	0.56	0, 3	0.70	66	0.69	0, 4	0.64	59
S8	Dual disorder treatment groups	0.31	0, 4	0.52	88	0.16	0, 2	0.55	91	0.28	0, 3	0.67	88
S9	Dual disorders model	0.56	0, 3	0.49	50	0.59	0, 1	0.48	41	0.53	0, 2	0.55	53
S10	Role of consumers on treatment team	0.22	0, 2	0.96	81	0.25	0, 4	0.91	88	0.28	0, 2	0.94	81
Nature of services mean		0.23	0, 1.0	0.76		0.18	0, 0.7	0.84		0.24	0, 0.7	0.79	
Total DACTS mean		0.14	0, 0.4	0.91		0.11	0, 0.5	0.91		0.17	0, 0.6	0.84	

reach our a priori cut-off of 0.80 for minimum intraclass correlation value. Intraclass correlations also indicated high agreement between phone and expert-scored self-report: 0.95, 0.86, 0.76, and 0.91 for the human resources, organizational boundaries, services subscale, and total DACTS score, respectively.

Similar to findings for consistency, inter-method consensus between onsite and remote measures was high for

total DACTS scores: 0.11 mean absolute difference for phone and 0.17 difference for expert-rated self-report. Phone subscale scores were also all within 0.18 of onsite scores, meeting our a priori expectations. The human resources subscale for expert-scored self-report was also within 0.17 of the onsite score for that subscale. However, similar to consistency measure findings, the discrepancy between expert-scored self-report and onsite scores for the organizational boundaries and services subscales, 0.26 and 0.24 points respectively, failed to meet a priori criteria of mean absolute differences of <0.2.

At the individual site level, the difference between total DACTS phone and on-site consensus scores was <0.08 for 17 (53 %) sites and within 0.11 for 22 (69 %) sites. One (3 %) outlier site exhibited an absolute difference of 0.5 between phone and on-site total DACTS scores. The difference between expert-scored self-report and on-site consensus scores for total DACTS was 0.25 or less for 26 (81 %) sites and within 0.11 for 15 (47 %) sites. Two (6 %) sites were outliers with absolute differences between expert-scored self-report and on-site total DACTS scores of 0.46 and 0.57.

At the item level, 16 items (57 %) rated via phone and 13 items (46 %) rated via expert-scored self-report had ICCs of 0.80 or above for inter-method consistency. For within-method consensus, only 11 items (39 %) rated via phone and 7 items (25 %) rated via expert-scored self-report had mean absolute differences <0.2. Percent agreement between raters was 75 % or higher for 15 (54 %) items using the phone method and 15 (54 %) items using expert-scored self-report. Lack of item variability in the sample adversely influenced some ICC values, artificially lowering or producing negative ICC values. For example, the intake rate item was rated universally high across all sites (score = 5), using all methods, with just a single discrepancy across the 32 teams where the self-report score was four at one site.

Repeated measures significance tests indicated no significant differences between methods for total DACTS score ($F(2, 29) = 2.34$, $p = 0.11$).

Discussion

The inter-rater reliability consistency for the alternative fidelity assessment methods was excellent for DACTS total scale scores and generally good for subscales. For example, the ICC for phone total DACTS inter-rater reliability (0.96) in this study exceeded the ICC found in an earlier study (McGrew et al. 2011) using ACT teams in a single state. In addition, the inter-rater reliability for both remote assessments was relatively close to the nearly perfect inter-rater reliability ICCs for onsite assessment demonstrated across 52 paired ratings in the National Evidence-Based Practices Project (ICC = 0.99) (McHugo et al. 2007). In contrast, at the item level, several items performed poorly in terms of inter-rater reliability and may require modification to improve remote methods when using item level versus subscale level scoring, such as explicit admission criteria, assertive engagement mechanisms, and dual disorders model. At least anecdotally, these items tend to involve subjective judgment to make ratings, even during on-site visits.

As expected, the phone method showed good concurrent validity with on-site total score and subscales when measured using both consistency and consensus measures. ICCs were between 0.84 (services subscale) and 0.96 (human resources subscale) for subscales and 0.91 for total DACTS, all slightly higher than a previous study (McGrew et al. 2011), with the exception of the services subscale. Mean absolute differences between phone and on-site scores also showed close consensus: 0.18 or less for all subscales and total DACTS. The difference in total DACTS scores, arguably the most important score for classifying programs as ACT (McGrew et al. 2011; McGrew, et al. 2013b; McHugo et al. 2007), between on-site and the remote assessments was only 0.11 (within 2.8 % of total range) for phone and 0.17 (within 4.3 % of total range) for expert-scored self-report, although the former was slightly higher than found in the earlier study (McGrew et al. 2011) (0.07). These differences still indicate relatively close consensus (>95 % accuracy), though perhaps not the level of equivalence required for the sole source for important policy distinctions, such as using phone-administered DACTS to qualify for special funding.

Although both consistency and consensus scores for expert-scored self-report were less favorable compared to phone method results, they were still sufficiently promising to continue work in this area. It should be noted that expert-scored self-report was the only method that allowed for no verification or communication from team leaders, therefore ambiguous information provided on tables could not be clarified as with other methods. This limitation is an artifact of the research protocol and would not be present in practice. As noted above, the onsite versus expert-scored self-report total DACTS mean absolute difference (0.17) was still <5 % of the range, and the incremental loss of consensus in moving from phone to expert-scored self-report methods is also relatively small and may represent an acceptable trade-off for lower burden, as one part of a hierarchical system of fidelity assessment methods (McGrew et al. 2011; McGrew, et al. 2013b).

Many of the problematic individual items also tended to reflect areas of the ACT model not commonly embraced by the MHICM program, such as full responsibility for services, 24-h crisis services, and dual disorders treatment

items. The DACTS was developed to measure adherence to ACT and may naturally suffer psychometrically when applied to non-ACT programs. As an example, MHICM teams were not intended to formally offer integrated dual disorders services. It should be noted, however, that smaller, generalist ACT teams also have tended not to offer integrated dual disorders services (McGrew and Bond 1995). Nevertheless, many sites were trying to address substance abuse needs, often informally or unsystematically, making accurate ratings difficult. Part of these issues could stem from somewhat ambiguous DACTS criteria for lower scores and/or our study raters having experience mostly rating teams who are required to score high on DACTS items to maintain funding (McGrew et al. 2011, 2013b). We attempted to further specify our DACTS protocol to clarify some of these scoring rules for the lower DACTS ratings of 1, 2, or 3 that were more commonly encountered in VA programs than in raters' previous experience. However, future research is needed to continue to add specificity to our protocol. Future work with a wider sample of teams scoring at both upper and lower ends of the scale will be needed to test these additions to see if item-level concurrent validity results for the remote methods improve. Interestingly, inter-rater reliability for integrated dual disorders treatment fidelity scale items also was the lowest (ICC = 0.89), among all five evidence-based practices assessed in the National EBP Project (McHugo et al. 2007), potentially indicating that these are difficult practices to rate reliably under a variety of circumstances.

Other problematic items in both phone and expert-scored self-report included three items derived from chart review: community based services, frequency of contact, and intensity of contact. Location of contact (facility or community) and service intensity were not systematically recorded by teams in our study and had to be estimated by the respondent in each condition, if the parameter was not available in the chart. To better understand the impact of real-world fidelity assessment challenges where optimal data are not always available, we re-examined our results by excluding these two items. Concurrent validity results did not change substantially. Inter-method ICCs were either the same or 0.01 better for the modified DACTS. Absolute mean differences between methods was 0.01 or 0.02 worse when excluding these items. Inter-rater ICCs for the modified DACTS did not change but absolute agreements between raters were actually worse when excluding these items, probably due to all methods experiencing the same data limitations. So even though our study setting presented challenges, likely similar to other real-world uses of remote fidelity assessments, the convergence of each method with on-site results was not compromised by these items. Certainly, scoring of these items will improve where location and length of service contacts are common and even required for Medicaid documentation. With the increasing availability of electronic medical records in mental health settings, remote record review is another option to explore in future work to improve upon remote fidelity assessment methods employed in this study.

We should note that we used fairly high criteria for excellent ICCs between methods (0.9 for inter-rater reliability and 0.8 for concurrent validity). Other classifications for ICCs in clinical assessments set lower standards, such as: <0.40 = poor; 0.40–0.59 = fair; 0.60–0.74 = good; and 0.75–1.00 = excellent (Cicchetti 1994). Using these criteria and excluding the intake item where there was almost perfect agreement and the ICC could not be calculated, most item-level inter-rater reliabilities would be classified as good, with only one phone item falling into the poor range (explicit admission criteria). In inter-method analyses of consistency with on-site scores, only one item for phone and expert-scored self-report (assertive engagement) would be classified as poor using the Cicchetti (1994) criteria (again, excluding the intake item).

Although our findings using less burdensome remote fidelity assessment is appealing, we advise caution in the wholesale replacement of on-site fidelity assessment. One key reason for caution is the inherent benefit of on-site evaluators providing technical assistance for program improvements during the course of on-site visits. Fidelity assessment is both a method of documenting adherence to a model but also a tool to provide specific feedback to reinforce strengths and improve areas of weakness. One idea for the use of remote fidelity assessment is to incorporate these methods in a stepped approach. For example, a system could include on-site fidelity assessments for a team's first year, followed by remote fidelity assessments. Further, periodic on-site assessment would be triggered by substantial program changes (e.g., key staff turnover), periods of low fidelity requiring close monitoring for follow-up, or other programmatic concerns. While further improvements could be made for remote fidelity assessments as noted above, the methods used in this study could be useful in large implementation efforts where on-site visits are cost prohibitive.

Limitations

To determine whether the phone and self-reported assessment methods made accurate classifications in situations that required a dichotomous judgment (for example, ACT vs non-ACT), sensitivity and specificity should be calculated. Unfortunately, this sample was limited to VA intensive case management teams that are intended to follow some, but not all, elements of assertive community treatment. Therefore, all teams scored below DACTS 4.0 total

average using each method and the average DACTS means across the sample was 3.2, closer to intensive case management scores than ACT (Salyers et al. 2003). This could cause concern that our study included only intensive case management rather than ACT services. However, given that ACT fidelity assessment is often used to distinguish high fidelity ACT from intensive case management and assign implementation fidelity scores for teams seeking to improve and become ACT-adherent, the reliability and validity demonstrated when using these teams is encouraging. Similarly, efforts to establish the psychometrics of the on-site DACTS also included MHICM teams (Teague et al. 1998; Salyers et al. 2003). Although it would have been ideal to have a range of ACT and intensive case management teams in a single study, this study did complement previous work (McGrew et al. 2011, 2013b) in that the current study used low scoring teams, previous work used high scoring teams, and all studies demonstrated promising results. However, further studies of alternative fidelity methods using a more diverse range of teams should report on the sensitivity and specificity.

Conclusions

Even using fairly stringent standards, both phone and expert-scored self-report fidelity assessment methods showed excellent inter-rater reliability, using both consensus and consistency measures, and excellent consistency and reasonably good absolute agreement with on-site total DACTS scores. Though our phone method did somewhat better than expert-scored self-report, both methods showed promise. More information regarding costs and team preferences for each method will be reported in subsequent manuscripts that should help to more sensitively weigh the pros and cons of the remote fidelity methods.

Acknowledgements This work was supported by VA Health Services Research and Development IIR 09-368.

References

Allness, D. J., & Knoedler, W. H. (1998). *The PACT model of community-based treatment for persons with severe and persistent mental illness: A manual for PACT start-up*. Arlington: NAMI.

Allness, D. J., & Knoedler, W. H. (2003). *The PACT model of community-based treatment for persons with severe and persistent mental illness: A manual for PACT start-up*. Arlington: NAMI.

Bond, G. R. (2013). Self-assessed fidelity: proceed with caution. *Psychiatric Services, 64*(4), 393–394. doi:10.1176/appi.ps.640418.

Bond, G. R., Drake, R. E., McHugo, G. J., Peterson, A. E., Jones, A. M., & Williams, J. (2014). Long-term sustainability of evidence-based practices in community mental health agencies. *Administration and Policy in Mental Health, 41*(2), 228–236. doi:10.1007/s10488-012-0461-5.

Bond, G. R., Drake, R. E., Mueser, K. T., & Latimer, E. (2001). Assertive community treatment for people with severe mental illness: Critical ingredients and impact on patients. *Disease Management & Health Outcomes, 9*, 141–159.

Bond, G. R., McGrew, J. H., & Fekete, D. M. (1995). Assertive outreach for frequent users of psychiatric hospitals: A meta-analysis. *Journal of Mental Health Administration, 22*(1), 4–16.

Bond, G. R., Williams, J., Evans, L., Salyers, M., Kim, H. W., Sharpe, H., & Leff, H. S. (2000). *Psychiatric rehabilitation fidelity toolkit*. Cambridge: Human Services Research Institute.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Clark, P. (2004). *The role of medicaid*, Chicago.

Drake, R. E., Bond, G. R., & Essock, S. M. (2009). Implementing evidence-based practices for people with schizophrenia. *Schizophrenia Bulletin, 35*(4), 704–713. doi:10.1093/schbul/sbp041.

Drake, R. E., Goldman, H. H., Leff, H. S., Lehman, A. F., Dixon, L., Mueser, K. T., & Torrey, W. C. (2001). Implementing evidence-based practices in routine mental health service settings. *Psychiatric Services, 52*(2), 179–182.

Herdelin, A. C., & Scott, D. L. (1999). Experimental studies of the program of assertive community treatment (PACT): A meta-analysis. *Journal of Disability Policy Studies, 10*, 53–89. doi:10.1177/104420739901000105.

Latimer, E. (1999). Economic impacts of assertive community treatment: A review of the literature. *Canadian Journal of Psychiatry, 44*, 443–454.

McGrew, J. H., & Bond, G. R. (1995). Critical ingredients of assertive community treatment: Judgments of the experts. *Journal of Mental Health Administration, 22*, 113–125.

McGrew, J. H., & Bond, G. R. (1997). The association between program characteristics and service delivery in assertive community treatment. *Administration and Policy in Mental Health and Mental Health Services, 25*, 175–189.

McGrew, J. H., Bond, G. R., Dietzen, L. L., & Salyers, M. P. (1994). Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology, 62*(4), 670–678.

McGrew, J. H., Stull, L. G., Rollins, A. L., Salyers, M. P., & Hicks, L. J. (2011). A comparison of phone-based and on-site assessment of fidelity for assertive community treatment in Indiana. *Psychiatric Services, 62*(6), 670–674. doi:10.1176/appi.ps.62.6.670.

McGrew, J. H., White, L. M., & Stull, L. G. (2013a). Self-assessed fidelity: Proceed with caution: In reply. *Psychiatric Services, 64*(4), 394. doi:10.1176/appi.ps.640419.

McGrew, J. H., White, L. M., Stull, L. G., & Wright-Berryman, J. (2013b). A comparison of self-reported and phone-administered methods of ACT fidelity assessment: A pilot study in Indiana. *Psychiatric Services, 64*(3), 272–276. doi:10.1176/appi.ps.001252012.

McHugo, G. J., Drake, R. E., Teague, G. B., & Xie, H. (1999). Fidelity to assertive community treatment and client outcomes in the new hampshire dual disorders study. *Psychiatric Services, 50*(6), 818–824.

McHugo, G. J., Drake, R. E., Whitley, R., Bond, G. R., Campbell, K., Rapp, C. A., & Finnerty, M. T. (2007). Fidelity outcomes in the national implementing evidence-based practices project. *Psychiatric Services, 58*(10), 1279–1284.

Mueser, K. T., Bond, G. R., Drake, R. E., & Resnick, S. G. (1998). Models of community care for severe mental illness: A review of

research on case management. *Schizophrenia Bulletin, 24*, 37–74.

Neale, M. S., Rosenheck, R., Castrodonatti, J., Martin, A., Morrissey, J., & D'Amico, M. (2007). *Mental health intensive case management (MHICM) in the department of veterans affairs: The tenth national performance monitoring report—FY 2006*. West Haven: VHA NEPEC Report.

Phillips, S. D., Burns, B. J., Edgar, E. R., Mueser, K. T., Linkins, K. W., Rosenheck, R. A., et al. (2001). Moving assertive community treatment into standard practice. *Psychiatric Services, 52*(6), 771–779.

President's New Freedom Commission on Mental Health. (2003). Achieving the promise: Transforming mental health care in America. Final Report. DHHS Pub. No. SMA-03-3832. Rockville: Substance Abuse and Mental Health Services Administration.

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine, 299*(17), 926–930.

Rosenheck, R. A., & Neale, M. S. (1998). Cost-effectiveness of intensive psychiatric community care for high users of inpatient services. *Archives of General Psychiatry, 55*(5), 459–466.

Rosenheck, R., Neale, M. S., Leaf, P., Milstein, R., & Frisman, L. (1995). Multisite experimental cost study of intensive psychiatric community care. *Schizophrenia Bulletin, 21*, 129–140.

Salyers, M. P., Bond, G. R., Teague, G. B., Cox, J. F., Smith, M. E., Hicks, M. L., & Koop, J. I. (2003). Is it ACT yet? Real-world examples of evaluating the degree of implementation for assertive community treatment. *Journal of Behavioral Health Services and Research, 30*(3), 304–320.

Salyers, M. P., McKasson, R. M., Bond, G. R., McGrew, J. H., Rollins, A. L., & Boyle, C. (2007). The role of technical assistance centers in implementing evidence-based practices: Lessons learned. *American Journal of Psychiatric Rehabilitation, 10*(2), 85–101.

Stein, L. I., & Test, M. A. (1980). An alternative to mental health treatment. I: Conceptual model, treatment program, and clinical evaluation. *Archives of General Psychiatry, 37*, 392–397.

Stemler, S. E. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation.(2004); *9*(4), 66–78. Retrieved January 2, 2015 from http://PAREonline.net/getvn.asp?v=9&n=4.

Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity in assertive community treatment: Development and use of a measure. *American Journal of Orthopsychiatry, 68*(2), 216–232.

Torrey, W. C., Finnerty, M., Evans, A., & Wyzik, P. F. (2003). Strategies for leading the implementation of evidence-based practices. *Psychiatric Clinics of North America, 26*, 883–897.

Ziguras, S., & Stuart, G. (2000). A meta-analysis of the effectiveness of mental health case management over 20 years. *Psychiatric Services, 51*, 1410–1421.