

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

Angela L. Rollins (alrollin@iupui.edu)

John H. McGrew (jmcgrew@iupui.edu)

Marina Kukla (mkukla@iupui.edu)

Alan B. McGuire (abmcguir@iupui.edu)

Mindy E. Flanagan (meflanag@iupui.edu)

Marcia G. Hunt (Marcia.hunt2@va.gov)

Doug L. Leslie (dleslie@phs.psu.edu)

Linda A. Collins (linda.collins4@va.gov)

Jennifer L. Wright-Berryman (jelmoore@iupui.edu)

Lia J. Hicks (lhicks@adultandchild.org)

Michelle P. Salyers (mpsalyer@iupui.edu)

Draft 9/27/14

Correspondence:

Angela L. Rollins
HSR&D Center for Health Information and Communication
Richard L. Roudebush VAMC
1481 W. 10th Street, 11-H
Indianapolis, IN 46202

Acknowledgements:

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

This is the author's manuscript of the article published in final edited form as:

Rollins, A. L., McGrew, J. H., Kukla, M., McGuire, A. B., Flanagan, M. E., Hunt, M. G., ... Salyers, M. P. (2015). Comparison of Assertive Community Treatment Fidelity Assessment Methods: Reliability and Validity. *Administration and Policy in Mental Health and Mental Health Services Research*, 1–11. <http://doi.org/10.1007/s10488-015-0641-1>

1
2
3
4 **Comparison of Assertive Community Treatment Fidelity Assessment Methods:**
5
6 **Reliability and Validity**

7
8
9 Abstract

10
11 Assertive community treatment is known for improving consumer outcomes, but is
12
13 difficult to implement. On-site fidelity measurement can help ensure model adherence, but is
14
15 costly in large systems. This study compared reliability and validity of three methods of fidelity
16
17 assessment (on-site, phone-administered, and expert-scored self-report) using a stratified
18
19 random sample of 32 mental health intensive case management teams from the Department of
20
21 Veterans Affairs (VA). Overall, phone, and to a lesser extent, expert-scored self-report fidelity
22
23 assessments compared favorably to onsite methods in inter-rater reliability and concurrent
24
25 validity. If used appropriately, these alternative protocols hold promise in monitoring large-scale
26
27 program fidelity with limited resources.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 **Comparison of Assertive Community Treatment Fidelity Assessment Methods:**
5
6 **Reliability and Validity**

7
8 Introduction
9

10 Assertive community treatment (ACT) is an effective model of community-based
11 treatment for people with severe mental illnesses (Stein & Test, 1980). ACT has been the
12 subject of over 30 randomized controlled trials. Typical outcomes include reduced hospital use,
13 increased housing stability, increased client retention, improvements in level of functioning and
14 quality of life, and increased satisfaction with treatment (Bond, Drake, Mueser, & Latimer, 2001;
15 Bond, McGrew, & Fekete, 1995; Herdelin & Scott, 1999; Mueser, Bond, Drake, & Resnick,
16 1998; Phillips et al., 2001; Ziguras & Stuart, 2000). In part because of the strong empirical
17 research base, ACT has gained broad acceptance and has been widely disseminated, both in
18 the United States and elsewhere. ACT has been identified as one of six evidence-based
19 practices for the public mental health sector (Drake et al., 2001), has been endorsed in U.S.
20 governmental reports (President's New Freedom Commission on Mental Health, 2003) and by
21 the U.S. Medicaid agency (Clark, 2004), and has received vigorous advocacy by the National
22 Alliance on Mental Illness (Allness & Knoedler, 1998, 2003; Torrey, Finnerty, Evans, & Wyzik,
23 2003). In addition, ACT is remarkable for the degree to which its structural and functional
24 features have been articulated (McGrew & Bond, 1995), as well as for having a widely-used
25 fidelity scale to assess a team's adherence to an ideal ACT model for staffing and services
26 (Teague, Bond, & Drake, 1998).
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

48 Although ACT programs are effective, they tend to be difficult to implement accurately.
49 Studies have found wide variability in the degree to which the self-designated "ACT" programs
50 adhere to the original design (McGrew & Bond, 1997; McGrew, Bond, Dietzen, & Salyers,
51 1994). Unfortunately, variability in implementation, as measured by departures from fidelity (i.e.,
52 degree of adherence to an intervention model), can critically affect outcomes in psychosocial
53 programs (Drake, Bond, & Essock, 2009; Latimer, 1999; McGrew et al., 1994; McHugo, Drake,
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 Teague, & Xie, 1999) and the sustainability of programs over time (G. R. Bond et al., 2014).
5
6 Given the fact that variability in program implementation is the norm and that uncontrolled
7
8 variability usually leads to poorer outcomes, there is now broad consensus on the need to verify
9
10 program fidelity for ACT and for other evidence-based practices. The current standard for
11
12 assessing ACT fidelity is the Dartmouth Assertive Community Treatment Scale (DACTS)
13
14 (Teague et al., 1998), which is administered by experts in the model through on-site visits to the
15
16 targeted program. The DACTS has good inter-rater reliability and can differentiate between
17
18 intensive case management models (Teague et al., 1998). A precursor of the DACTS showed a
19
20 robust correlation between ACT fidelity and reduced hospital use (McGrew et al., 1994), which
21
22 justifies efforts on the part of health care systems to assess fidelity.
23
24
25

26
27 Although there is general agreement on the need for fidelity assessment, there is
28
29 disagreement about the reliability, validity and cost effectiveness (e.g., assessment burden) of
30
31 different assessment methods (Bond, 2013; McGrew, White, & Stull, 2013; McGrew, White,
32
33 Stull, & Wright-Berryman, 2013). Due to state and federal fiscal restraints, gold standard on-site
34
35 visits may not be feasible within large health systems due to their cost in personnel, time, and
36
37 lost productivity for clinicians. This dilemma has led some to propose alternative methods, such
38
39 as phone-based and more innovative self-report methods where data are provided by the team
40
41 but scoring is still done by an expert rater (“expert-scored self-report”). Two studies examined
42
43 these alternate methods for ACT. In the first study, McGrew and colleagues (2011) showed that
44
45 phone-administered DACTS fidelity could be rated reliably and had good agreement with on-site
46
47 assessment as measured by between-method consistency (i.e., inter-rater reliability) and
48
49 consensus (i.e., low mean absolute differences in scores). In a second study, they
50
51 demonstrated that expert-scored self-reported fidelity could be rated reliably and was
52
53 comparable to phone-administered fidelity, again, as indicated by good consistency and
54
55 consensus (McGrew, White, Stull, et al., 2013). However, both studies were small, used
56
57 convenience sampling, took place in a single state and were limited to well-established, stable
58
59
60
61
62
63
64
65

1
2
3
4 teams with both good overall fidelity and extensive prior experience in fidelity assessments.
5
6 These limitations in external validity are problematic because fidelity monitoring, arguably, is
7
8 equally if not more critical for developing or relatively new teams and for teams that do not have
9
10 a history of good fidelity. Before such a substantial and economically appealing change to
11
12 current fidelity assessment methods can be recommended, remote fidelity methods research
13
14 requires a strong replication in another health system, using a more rigorous, prospective
15
16 design and randomized sampling of teams with a wider range of ACT fidelity. The current study
17
18 addresses these needs in applying the alternative fidelity methods in a larger, nationwide study
19
20 using a rigorous design and sampling framework where uniformly high fidelity to ACT was not
21
22 expected. Moreover, conclusions about the validity of expert-scored self-report fidelity are
23
24 problematic because the comparison was to phone-administered fidelity, not on-site fidelity,
25
26 which is still considered the gold standard. The current study will allow us to directly test the
27
28 comparability of expert-scored self- report with on-site methods using a prospective study
29
30 design and blinded raters in each condition.
31
32
33
34

35 The primary aims of the current study were to examine the reliability and concurrent
36
37 validity of three different methods of fidelity assessment through a comparison of expert-scored
38
39 self-report, phone, and on-site fidelity assessment. Based on high DACTS inter-rater reliability
40
41 for on-site assessment (ICC=.99) (McHugo et al., 1999) and previous work with phone methods,
42
43 (McGrew et al., 2011; McGrew, White, Stull, et al., 2013), we expected inter-rater reliability
44
45 (consistency) above .9 and inter-rater consensus as indicated by mean absolute differences of
46
47 less than .1 (2.5% of the scoring range). Based on promising results from the prior published
48
49 work (McGrew et al., 2011), we also expected both remote methods to have excellent
50
51 consistency with on-site results as indicated by ICCs of at least .80 and good consensus with
52
53 on-site results as indicated by mean absolute differences of less than .2 (5% of the scoring
54
55 range). Item-level results were also explored.
56
57
58
59
60
61
62
63
64
65

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, Neale, Leaf, Milstein, & Frisman, 1995; Rosenheck & 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 2011-2013. The pool of potential sites was limited to MHICM teams in existence for one 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

1
2
3
4 prevent “spectrum bias” – that a test may have differential predictive validity if tested in an
5
6 extreme group (Ransohoff & Feinstein, 1978) – we also stratified based on team prior year self-
7
8 reported fidelity scores using overall sample median split. This resulted in 4 strata: high and low
9
10 fidelity GM&S and high and low fidelity NP sites. Teams were selected randomly within each of
11
12 the 4 strata. Ten high and 10 low GM&S sites and 6 high and 6 low NP sites were invited to
13
14 participate. When a site declined to participate, we replaced it with the next site on our list.
15
16

17
18 As part of recruitment efforts, we distributed brochures and made presentations to
19
20 national MHICM teleconferences to inform MHICM team leaders regarding the study. All
21
22 procedures were reviewed and approved by the VA’s Central IRB and [local] VA Medical
23
24 Center’s Research and Development Committee.
25
26

27 28 Measures

29
30 Dartmouth Assertive Community Treatment Scale (DACTS). The DACTS (Teague et al.,
31
32 1998) is a 28-item scale that assesses degree of fidelity to the ACT model along three
33
34 dimensions: Human Resources (e.g., small caseload, psychiatrist on staff), Organizational
35
36 Boundaries (e.g., explicit admission criteria), and Nature of Services (e.g., in-vivo services).
37
38 Each item is rated on a 5-point behaviorally anchored scale, ranging from 1 = not implemented
39
40 to 5 = fully implemented. Anchors are item-specific. The DACTS has been shown to
41
42 discriminate between four types of services (Teague et al., 1998) and is sensitive to change
43
44 over time in implementation efforts (McHugo et al., 2007). Inter-rater reliability of the on-site
45
46 DACTS (between two trained raters making a conjoint visit) was found to be .99 in the National
47
48 Implementing Evidence-based Practices Project (McHugo et al., 2007). All three forms of fidelity
49
50 assessment were based on the DACTS, but used different methods for data collection and
51
52 rating as described below.
53
54
55
56
57
58
59

60 Procedures

61
62
63
64
65

1
2
3
4 We counterbalanced the order of phone and on-site fidelity assessments, such that half
5
6 the sites were randomly assigned to receive the phone assessment first, and the other half to
7
8 receive the on-site assessment first. However, because three sites originally assigned to
9
10 receive phone first rescheduled so that on-site came first, 19 sites received on-site assessments
11
12 first and 13 sites received the phone assessment first. After agreeing to participate, each
13
14 MHICM program leader and his/her supervisor were contacted via email and/or phone to begin
15
16 preparations for their first assessment. Team leaders prepared a set of 10 tables describing
17
18 objective team composition and activities (e.g., admission criteria, number of veterans receiving
19
20 services offered, description of recent circumstances around hospitalizations) in advance of the
21
22 assessment. These tables were used for all 3 assessments and were expanded from previously
23
24 established fidelity protocols and from those used in prior studies by McGrew and colleagues
25
26 (2011). Revisions included more comprehensive instructions for completing table items (e.g.,
27
28 more precise definitions of graduation or dropout from ACT, rephrasing questions on team
29
30 meetings that capture the full array of potential content and frequency that fall below higher
31
32 DACTS ratings), adding table items to more fully assess team admission criteria and
33
34 procedures and assertive engagement examples, and adding VA-specific terminology to avoid
35
36 confusion for the respondent. Expert-scored self-report was completed solely from these tables.
37
38 Phone and On-site assessments used the tables as the basis for the interviews with the team
39
40 leader and others. Specific procedures for each method are outlined below. Raters for each
41
42 assessment type were blinded to the results of other assessment types for that site.
43
44
45
46
47

48
49 On-site fidelity assessment. Prior to the visit, the MHICM team leader received a
50
51 checklist of items and data collection sources needed for the on-site visit (e.g., team roster,
52
53 chart reviews, interviews with specific staff members). On-site visit days were scheduled so that
54
55 the assessor, the MHICM program leader, and as many key MHICM staff members as possible
56
57 were present for the visit. Each day-long on-site visit included observation of the daily team
58
59 meeting, interviews with the program leader, vocational, peer, and substance abuse specialists
60
61
62
63
64
65

1
2
3
4 (if assigned to the team), shadowing team members in the community, and reviewing a random
5
6 sample of charts and other records. On-site DACTS fidelity assessments used a single rater
7
8 because inter-rater reliability has been determined to be strong (Salyers et al., 2007). However,
9
10 to ensure quality control, the first author attended 4 of the 32 on-site assessments to monitor
11
12 inter-rater reliability of the on-site approach. The paired ratings for these sites averaged a
13
14 difference of .03 on the total DACTS score and an ICC of .99, indicating very high inter-rater
15
16 consensus and consistency, consistent with previous findings (McHugo et al., 2007).
17
18

19
20 Phone fidelity assessment. Similar to the on-site visits, team leaders completed the
21
22 fidelity tables in advance. As part of this effort, sites were asked to report de-identified service
23
24 use data from electronic health records for a random sample of charts. A phone assessor was
25
26 available to address questions about preparing the fidelity tables or accessing health records
27
28 before the interview (e.g., suggestions about where to find data for the tables, clarify
29
30 terminology, etc.). Team leaders were encouraged to fax or email completed tables to the
31
32 program manager in advance of the phone fidelity call to streamline the phone interview
33
34 process. To test the validity of phone interviews when conducted with the least possible burden
35
36 to program staff, we only required the team leader's participation. In addition, team leaders
37
38 should provide more accurate information than other team members (Bond, Williams, Evans,
39
40 Salyers, Kim, Sharpe, & Leff, 2000). Two raters conducted the phone assessments via
41
42 conference call, independently scored the items on the fidelity scale, and later came to
43
44 consensus on discrepant items.
45
46
47

48
49 Expert-scored self-report fidelity assessment. This assessment method used the tables
50
51 prepared by teams in advance of on-site and phone fidelity as data to score each DACTS item.
52
53 Two fidelity assessors used the prepared tables without any contact with or clarifications from
54
55 respondents, independently scored the items, and later came to consensus on discrepant items.
56
57

58 Fidelity Assessors. On-site assessments were conducted by one of two raters who
59
60 alternated between performing the on-site or phone-based assessment. Three additional raters
61
62
63
64
65

1
2
3
4 rotated roles as the second phone-based fidelity rater and one of the two experts scoring self-
5 reported data. All raters were experienced fidelity assessors and participated in a day-long
6 workshop to review DACTS fidelity assessment materials adapted to the VA context. The raters
7 also attended monthly calls throughout the project to discuss scoring rules and protocols for
8 ongoing quality assurance. Fidelity data were entered into a relational database that included
9 embedded computation and scoring capabilities to reduce mathematical errors (e.g., calculation
10 of caseload ratio).
11
12
13
14
15
16
17
18
19
20
21

22 Data Analysis

23
24 Descriptive statistics were computed for on-site fidelity data to give an overall sense of
25 how close the MHICM sample adhered to ACT standards. Three indicators were used to assess
26 inter-rater agreement (reliability) and inter-method agreement (concurrent validity): consistency,
27 calculated with the two-way mixed intraclass correlation coefficient (ICC); consensus, estimated
28 from the mean of the absolute value of the difference between raters or methods; and percent
29 agreement (Stemler, 2004). We examined inter-rater reliability, i.e., consistency and consensus,
30 for the total score and for each subscale of the DACTS for both phone and expert-scored self-
31 report methods. Concurrent validity, i.e., consistency and consensus, was calculated between
32 all three measures of fidelity for the corresponding DACTS total and subscale scores. Mean
33 differences between the three assessments also were tested using a mixed model (repeated
34 measures) approach, followed by Tukey's pairwise comparisons to adjust for multiple
35 comparisons.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53 Results

54
55 As seen in Table 1, teams showed modest fidelity to the ACT model as assessed by the
56 on-site method. DACTS score means were below 4.0 (fully implemented) for all subscales and
57
58
59
60
61
62
63
64
65

1
2
3
4 the total scale score. No team in our sample scored a 4.0 or higher on the Total DACTS mean
5
6 using any fidelity assessment method.
7
8
9

10 Inter-rater reliability

11
12
13 At the DACTS subscale and total scale level, analyses indicated good inter-rater
14
15 reliability for phone and expert-scored self-report methods using both consistency and
16
17 consensus measures. As seen in Table 2, consistency, as measured using inter-rater reliability
18
19 (intraclass correlations) was very good (ICCs=.96) for total DACTS scores for both phone and
20
21 expert-scored self-report and also was high for subscales for each alternative method, with the
22
23 lowest ICC for the Services subscale for the phone method (.81). Likewise, consensus was
24
25 high for subscales and total DACTS mean scores with all mean absolute differences falling
26
27 below .20. The mean absolute difference between raters on the total DACTS score was .09
28
29 and .10 for phone and self-report assessment methods, respectively, indicating high consensus
30
31 between raters for both remote methods.
32
33
34

35 At the item level, 23 items (82%) rated via phone and 22 items (79%) rated via expert-
36
37 scored self-report had ICCs of .80 or above for inter-rater consistency. For consensus, 17 of the
38
39 28 (61%) items using the phone method and 12 of 28 items (43%) using expert-scored self-
40
41 report had mean absolute differences less than .2, indicating high levels of consensus between
42
43 raters on those items. Percent agreement between raters was 75% or higher for 23 (82%)
44
45 items using the phone method and 19 (68%) items using expert-scored self-report.
46
47
48
49
50

51 Concurrent validity

52
53 Inter-method agreement between phone, expert-scored self-report, and on-site methods
54
55 was high for total DACTS score and most subscales using consistency measures. As seen in
56
57 Table 3, intraclass correlations between phone and on-site methods were .96, .85, .84, and .91
58
59 for the Human Resources, Organizational Boundaries, Services subscale, and total DACTS
60
61
62
63
64
65

1
2
3
4 score, respectively (all above the .80 criterion for phone). Intraclass correlations indicating
5 agreement between self-report and on-site methods were slightly lower: .92, .66, .79, and .84
6
7 for the Human Resources, Organizational Boundaries, Services subscale, and total DACTS
8
9 score, respectively. The Organizational boundaries and Services subscales for expert-scored
10
11 self-report were the only subscales that did not reach our *a priori* cut-off of .80 for minimum
12
13 intraclass correlation value. Intraclass correlations also indicated high agreement between
14
15 phone and expert-scored self-report: .95, .86, .76, and .91 for the Human Resources,
16
17 Organizational Boundaries, Services subscale, and total DACTS score, respectively.
18
19
20
21

22 Similar to findings for consistency, inter-method consensus between onsite and remote
23
24 measures was high for total DACTS scores: .11 mean absolute difference for phone and .17
25
26 difference for expert-rated self-report. Phone subscale scores were also all within .18 of onsite
27
28 scores, meeting our *a priori* expectations. The Human resources subscale for expert-scored
29
30 self-report was also within .17 of the onsite score for that subscale. However, similar to
31
32 consistency measure findings, the discrepancy between expert-scored self-report and onsite
33
34 scores for the Organizational boundaries and Services subscales, .26 and .24 points
35
36 respectively, failed to meet *a priori* criteria of mean absolute differences of less than .2.
37
38
39

40 At the individual site level, the difference between total DACTS phone and on-site
41
42 consensus scores was less than .08 for 17 (53%) sites and within .11 for 22 (69%) sites. One
43
44 (3%) outlier site exhibited an absolute difference of .5 between phone and on-site total DACTS
45
46 scores. The difference between expert-scored self-report and on-site consensus scores for total
47
48 DACTS was .25 or less for 26 (81%) sites and within .11 for 15 (47%) sites. Two (6%) sites
49
50 were outliers with absolute differences between expert-scored self-report and on-site total
51
52 DACTS scores of .46 and .57.
53
54

55 At the item level, 16 items (57%) rated via phone and 13 items (46%) rated via expert-
56
57 scored self-report had ICCs of .80 or above for inter-method consistency. For within-method
58
59 consensus, only 11 items (39%) rated via phone and 7 items (25%) rated via expert-scored self-
60
61
62
63
64
65

1
2
3
4 report had mean absolute differences less than .2. Percent agreement between raters was 75%
5
6 or higher for 15 (54%) items using the phone method and 15 (54%) items using expert-scored
7
8 self-report. Lack of item variability in the sample adversely influenced some ICC values,
9
10 artificially lowering or producing negative ICC values. For example, the intake rate item was
11
12 rated universally high across all sites (score=5), using all methods, with just a single
13
14 discrepancy across the 32 teams where the self-report score was 4 at one site.
15
16

17 Repeated measures significance tests indicated no significant differences between
18
19 methods for total DACTS score ($F(2, 29)=2.34, p=0.11$).
20
21
22

23 Discussion

24
25
26 The inter-rater reliability consistency for the alternative fidelity assessment methods was
27
28 excellent for DACTS total scale scores and generally good for subscales. For example, the ICC
29
30 for phone total DACTS inter-rater reliability (.96) in this study exceeded the ICC found in an
31
32 earlier study (McGrew et al., 2011) using ACT teams in a single state. In addition, the inter-rater
33
34 reliability for both remote assessments was relatively close to the nearly perfect inter-rater
35
36 reliability ICCs for onsite assessment demonstrated across 52 paired ratings in the National
37
38 Evidence-Based Practices Project (ICC=.99) (McHugo et al., 2007). In contrast, at the item
39
40 level, several items performed poorly in terms of inter-rater reliability and may require
41
42 modification to improve remote methods when using item level vs. subscale level scoring, such
43
44 as explicit admission criteria, assertive engagement mechanisms, and dual disorders model. At
45
46 least anecdotally, these items tend to involve subjective judgment to make ratings, even during
47
48 on-site visits.
49
50
51

52
53 As expected, the phone method showed good concurrent validity with on-site total score
54
55 and subscales when measured using both consistency and consensus measures. ICCs were
56
57 between .84 (Services subscale) and .96 (Human Resources subscale) for subscales and .91
58
59 for total DACTS, all slightly higher than a previous study (McGrew et al., 2011), with the
60
61
62
63
64
65

1
2
3
4 exception of the Services subscale. Mean absolute differences between phone and on-site
5
6 scores also showed close consensus: .18 or less for all subscales and total DACTS. The
7
8 difference in total DACTS scores, arguably the most important score for classifying programs as
9
10 ACT (McGrew et al., 2011; McGrew, White, Stull, et al., 2013; McHugo et al., 2007), between
11
12 on-site and the remote assessments was only .11 (within 2.8% of total range) for phone and .17
13
14 (within 4.3% of total range) for expert-scored self-report, although the former was slightly higher
15
16 than found in the earlier study (McGrew et al., 2011) (.07).. These differences still indicate
17
18 relatively close consensus (greater than 95% accuracy), though perhaps not the level of
19
20 equivalence required for the sole source for important policy distinctions, such as using phone-
21
22 administered DACTS to qualify for special funding.
23
24
25

26
27 Although both consistency and consensus scores for expert-scored self-report were less
28
29 favorable compared to phone method results, they were still sufficiently promising to continue
30
31 work in this area. It should be noted that expert-scored self-report was the only method that
32
33 allowed for no verification or communication from team leaders, therefore ambiguous
34
35 information provided on tables could not be clarified as with other methods. This limitation is an
36
37 artifact of the research protocol and would not be present in practice. As noted above, the onsite
38
39 vs. expert-scored self-report total DACTS mean absolute difference (.17) was still less than 5%
40
41 of the range, and the incremental loss of consensus in moving from phone to expert-scored self-
42
43 report methods is also relatively small and may represent an acceptable trade-off for lower
44
45 burden, as one part of a hierarchical system of fidelity assessment methods (McGrew et al.,
46
47 2011; McGrew, White, Stull, et al., 2013).
48
49
50

51
52 Many of the problematic individual items also tended to reflect areas of the ACT model
53
54 not commonly embraced by the MHICM program, such as full responsibility for services, 24-
55
56 hour crisis services, and dual disorders treatment items. The DACTS was developed to
57
58 measure adherence to ACT and may naturally suffer psychometrically when applied to non-ACT
59
60 programs. As an example, MHICM teams were not intended to formally offer integrated dual
61
62
63
64
65

1
2
3
4 disorders services. It should be noted, however, that smaller, generalist ACT teams also have
5
6 tended not to offer integrated dual disorders services (McGrew and Bond, 1995). Nevertheless,
7
8 many sites were trying to address substance abuse needs, often informally or unsystematically,
9
10 making accurate ratings difficult. Part of these issues could stem from somewhat ambiguous
11
12 DACTS criteria for lower scores and/or our study raters having experience mostly rating teams
13
14 who are required to score high on DACTS items to maintain funding (McGrew et al., 2011;
15
16 McGrew, White, Stull, et al., 2013). We attempted to further specify our DACTS protocol to
17
18 clarify some of these scoring rules for the lower DACTS ratings of 1, 2, or 3 that were more
19
20 commonly encountered in VA programs than in raters' previous experience. However, future
21
22 research is needed to continue to add specificity to our protocol. Future work with a wider
23
24 sample of teams scoring at both upper and lower ends of the scale will be needed to test these
25
26 additions to see if item-level concurrent validity results for the remote methods improve.
27
28 Interestingly, inter-rater reliability for integrated dual disorders treatment fidelity scale items also
29
30 was the lowest (ICC=.89), among all 5 evidence-based practices assessed in the National EBP
31
32 Project (McHugo et al., 2007), potentially indicating that these are difficult practices to rate
33
34 reliably under a variety of circumstances.
35
36
37
38
39

40 Other problematic items in both phone and expert-scored self-report included three
41
42 items derived from chart review: community based services, frequency of contact, and intensity
43
44 of contact. Location of contact (facility or community) and service intensity were not
45
46 systematically recorded by teams in our study and had to be estimated by the respondent in
47
48 each condition, if the parameter was not available in the chart. To better understand the impact
49
50 of real-world fidelity assessment challenges where optimal data are not always available, we re-
51
52 examined our results by excluding these two items. Concurrent validity results did not change
53
54 substantially. Inter-method ICCs were either the same or .01 better for the modified DACTS.
55
56 Absolute mean differences between methods was .01 or .02 worse when excluding these items.
57
58 Inter-rater ICCs for the modified DACTS did not change but absolute agreements between
59
60
61
62
63
64
65

1
2
3
4 raters were actually worse when excluding these items, probably due to all methods
5
6 experiencing the same data limitations. So even though our study setting presented challenges,
7
8 likely similar to other real-world uses of remote fidelity assessments, the convergence of each
9
10 method with on-site results was not compromised by these items. Certainly, scoring of these
11
12 items will improve where location and length of service contacts are common and even required
13
14 for Medicaid documentation. With the increasing availability of electronic medical records in
15
16 mental health settings, remote record review is another option to explore in future work to
17
18 improve upon remote fidelity assessment methods employed in this study.
19
20
21

22 We should note that we used fairly high criteria for excellent ICCs between methods (.9
23
24 for inter-rater reliability and .8 for concurrent validity). Other classifications for ICCs in clinical
25
26 assessments set lower standards, such as: under .40 = poor; .40 - .59 = fair; .60 - .74 = good;
27
28 and .75 – 1.00 = excellent (Cicchetti, 1994). Using these criteria and excluding the intake item
29
30 where there was almost perfect agreement and the ICC could not be calculated, most item-level
31
32 inter-rater reliabilities would be classified as good, with only one phone item falling into the poor
33
34 range (explicit admission criteria). In inter-method analyses of consistency with on-site scores,
35
36 only one item for phone and expert-scored self-report (assertive engagement) would be
37
38 classified as poor using the Cicchetti (1994) criteria (again, excluding the Intake item).
39
40
41

42 Although our findings using less burdensome remote fidelity assessment is appealing,
43
44 we advise caution in the wholesale replacement of on-site fidelity assessment. One key reason
45
46 for caution is the inherent benefit of on-site evaluators providing technical assistance for
47
48 program improvements during the course of on-site visits. Fidelity assessment is both a method
49
50 of documenting adherence to a model but also a tool to provide specific feedback to reinforce
51
52 strengths and improve areas of weakness. One idea for the use of remote fidelity assessment is
53
54 to incorporate these methods in a stepped approach. For example, a system could include on-
55
56 site fidelity assessments for a team's first year, followed by remote fidelity assessments.
57
58 Further, periodic on-site assessment would be triggered by substantial program changes (e.g.,
59
60
61
62
63
64
65

1
2
3
4 key staff turnover), periods of low fidelity requiring close monitoring for follow-up, or other
5
6 programmatic concerns. While further improvements could be made for remote fidelity
7
8 assessments as noted above, the methods used in this study could be useful in large
9
10 implementation efforts where on-site visits are cost prohibitive.
11
12
13

14 15 Limitations

16
17 To determine whether the phone and self-reported assessment methods made accurate
18
19 classifications in situations that required a dichotomous judgment (for example, ACT versus
20
21 non-ACT), sensitivity and specificity should be calculated. Unfortunately, this sample was limited
22
23 to VA intensive case management teams that are intended to follow some, but not all, elements
24
25 of assertive community treatment. Therefore, all teams scored below DACTS 4.0 total average
26
27 using each method and the average DACTS means across the sample was 3.2, closer to
28
29 intensive case management scores than ACT (Salyers, Bond, Teague, Cox, Smith, Hicks, &
30
31 Koop, 2003). This could cause concern that our study included only intensive case
32
33 management rather than ACT services. However, given that ACT fidelity assessment is often
34
35 used to distinguish high fidelity ACT from intensive case management and assign
36
37 implementation fidelity scores for teams seeking to improve and become ACT-adherent, the
38
39 reliability and validity demonstrated when using these teams is encouraging. Similarly, efforts to
40
41 establish the psychometrics of the on-site DACTS also included MHICM teams (Teague et al.,
42
43 1998; Salyers et al., 2003). Although it would have been ideal to have a range of ACT and
44
45 intensive case management teams in a single study, this study did complement previous work
46
47 (McGrew et al., 2011; McGrew, White, Stull, et al., 2013) in that the current study used low
48
49 scoring teams, previous work used high scoring teams, and all studies demonstrated promising
50
51 results. However, further studies of alternative fidelity methods using a more diverse range of
52
53 teams should report on the sensitivity and specificity.
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 Conclusions
5

6 Even using fairly stringent standards, both phone and expert-scored self-report fidelity
7 assessment methods showed excellent inter-rater reliability, using both consensus and
8 consistency measures, and excellent consistency and reasonably good absolute agreement
9 with on-site total DACTS scores. Though our phone method did somewhat better than expert-
10 scored self-report, both methods showed promise. More information regarding costs and team
11 preferences for each method will be reported in subsequent manuscripts that should help to
12 more sensitively weigh the pros and cons of the remote fidelity methods.
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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

Table 2. Inter-rater reliability indicators for phone and expert-scored self-report methods: Intra-class correlations (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	.98	0.03	96.9	.71	0.25	81.3
H2	Team approach	1.0	0	100.0	1.0	0.03	96.9
H3	Program meeting	.97	0.22	78.1	.83	0.43	71.9
H4	Practicing team leader	.96	0.22	81.3	.91	0.47	59.4
H5	Continuity of staffing	.87	0.34	75.0	.93	0.25	75.0
H6	Staff capacity	.88	0.22	81.3	.82	0.25	75.0
H7	Psychiatrist on staff	.97	0.16	87.5	.90	0.28	87.5
H8	Nurse on Staff	.96	0.03	96.9	.91	0.09	93.8
H9	Substance Abuse Specialist on Staff	.95	0.09	93.8	1.0	0	100
H10	Vocational Specialist on Staff	.97	0.06	93.8	.74	0.16	87.5
H11	Program Size	.99	0.06	93.8	.94	0.22	81.3
	Human Resources Mean	.96	0.09		.92	0.13	
O1	Explicit Admission Criteria	.38	0.56	46.9	.60	0.50	56.3
O2	Intake Rate	1.0	0	100	0	0	93.8
O3	Full Responsibility for Treatment Services	.73	0.56	46.9	.77	0.53	56.3
O4	Responsibility for Crisis Services	.90	0.41	75.0	.95	0.38	62.5
O5	Responsibility for Hospital Admissions	.87	0.25	78.1	.88	0.34	65.6
O6	Responsibility for Hospital Discharge Planning	.89	0.19	84.4	.92	0.19	84.4
O7	Time-Unlimited Services	.97	0.03	96.9	.90	0.13	87.5
	Organizational Boundaries Mean	.81	0.18		.87	0.16	
S1	In-vivo services	1.0	0	100	1.0	0	100
S2	No drop out policy	.94	0.06	93.8	.77	0.19	81.3
S3	Assertive engagement mechanisms	.52	0.63	53.1	.49	0.75	43.8
S4	Intensity of service	.93	0.09	93.8	.94	0.09	93.8
S5	Frequency of contact	.88	0.06	93.8	.96	0.06	93.8
S6	Work with support system	.96	0.16	87.5	.98	0.13	84.4
S7	Individualized substance abuse treatment	.86	0.38	68.8	.86	0.47	62.5
S8	Dual disorder treatment groups	.70	0.19	90.6	.98	0.06	93.8
S9	Dual disorders model	.60	0.50	53.1	.81	0.56	46.9
S10	Role of consumers on treatment team	.97	0.19	84.4	.88	0.28	84.4
	Nature of Services Mean	.88	0.16		.91	0.19	
	Total DACTS Mean	.96	0.09		.96	0.10	

Table 3. Intermethod agreement using mean absolute differences, range of absolute differences, intra-class correlations, 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	.86	84	0.13	0, 1	.89	88	0.16	0, 1	.86	84
H2	Team approach	0.13	0, 2	.96	94	0.53	0, 2	.88	53	0.53	0, 2	.89	53
H3	Program meeting	0.50	0, 4	.81	69	0.63	0, 2	.85	53	0.56	0, 4	.79	66
H4	Practicing team leader	0.28	0, 2	.93	78	0.31	0, 2	.94	72	0.34	0, 2	.91	75
H5	Continuity of staffing	0.31	0, 2	.89	72	0.16	0, 1	.96	84	0.34	0, 2	.88	69
H6	Staff capacity	0.34	0, 2	.65	72	0.19	0, 1	.90	81	0.34	0, 3	.55	75
H7	Psychiatrist on staff	0.28	0, 3	.92	84	0.19	0, 2	.95	88	0.16	0, 3	.95	91
H8	Nurse on Staff	0.09	0, 3	.73	97	0.13	0, 2	.77	91	0.22	0, 3	.54	88
H9	Substance Abuse Specialist on Staff	0.09	0, 2	.95	94	0.09	0, 2	.95	94	0.06	0, 2	.96	97
H10	Vocational Specialist on Staff	0.22	0, 3	.59	88	0.28	0, 4	.35	88	0.19	0, 3	.61	88
H11	Program Size	0.13	0, 1	.97	88	0.13	0, 1	.97	88	0.25	0, 1	.95	75
	Human Resources Mean	0.14	0, 0.5	.95		0.14	0, 0.4	.96		0.17	0, 0.5	.92	
O1	Explicit Admission Criteria	0.59	0, 2	.57	47	0.50	0, 2	.52	56	0.53	0, 3	.55	56
O2	Intake Rate	0.06	0, 1	-.07	91	0.06	0, 1	-.07	94	0.06	0, 1	-.07	91
O3	Full Responsibility for Treatment Services	0.66	0, 2	.66	50	0.66	0, 2	.67	44	0.75	0, 3	.47	47
O4	Responsibility for Crisis Services	0.41	0, 2	.92	66	0.47	0, 2	.90	69	0.44	0, 4	.89	72
O5	Responsibility for Hospital Admissions	0.44	0, 2	.75	63	0.31	0, 1	.89	69	0.31	0, 2	.88	72
O6	Responsibility for Hospital Discharge Planning	0.47	0, 3	.64	66	0.25	0, 2	.83	78	0.28	0,2	.83	78

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

O7	Time-Unlimited Services	0.06	0, 2	.90	97	0.09	0, 1	.92	91	0.09	0,1	.93	91
	Organizational Boundaries Mean	0.20	0, 0.7	.86		0.18	0, 0.9	.85		0.26	0, 1.3	.66	
S1	In-vivo services	0.09	0, 1	.95	91	0.31	0, 2	.56	78	0.34	0, 3	.51	78
S2	No drop out policy	0.09	0, 1	.90	91	0.16	0, 1	.82	84	0.19	0, 1	.78	81
S3	Assertive engagement mechanisms	0.59	0, 2	.63	53	0.63	0, 2	.46	50	0.84	0, 2	.32	41
S4	Intensity of service	0.19	0, 3	.82	88	0.47	0, 2	.64	59	0.53	0, 3	.43	63
S5	Frequency of contact	0.09	0, 3	.67	97	0.28	0, 1	.59	72	0.38	0, 3	.44	69
S6	Work with support system	0.47	0, 2	.88	66	0.66	0, 3	.80	50	0.44	0, 2	.90	66
S7	Individualized substance abuse treatment	0.56	0, 3	.74	59	0.56	0, 3	.70	66	0.69	0, 4	.64	59
S8	Dual disorder treatment groups	0.31	0, 4	.52	88	0.16	0, 2	.55	91	0.28	0, 3	.67	88
S9	Dual disorders model	0.56	0, 3	.49	50	0.59	0, 1	.48	41	0.53	0, 2	.55	53
S10	Role of consumers on treatment team	0.22	0, 2	.96	81	0.25	0, 4	.91	88	0.28	0, 2	.94	81
	Nature of Services Mean	0.23	0, 1.0	.76		0.18	0, 0.7	.84		0.24	0, 0.7	.79	
	Total DACTS Mean	0.14	0, 0.4	.91		0.11	0, 0.5	.91		0.17	0, 0.6	.84	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

References

Allness, D. J., & Knodler, 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, VA: NAMI.

Allness, D. J., & Knodler, 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, VA: 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. *Adm Policy Ment 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, MA: Human Services Research

Institute.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology (Vol. 6, pp. 284-290). US: American Psychological Association.

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

- 1
2
3
4 Drake, R. E., Bond, G. R., & Essock, S. M. (2009). Implementing evidence-based practices for
5
6 people with schizophrenia. *Schizophrenia Bulletin*, 35(4), 704-713. doi:
7
8 10.1093/schbul/sbp041
9
- 10 Drake, R. E., Goldman, H. H., Leff, H. S., Lehman, A. F., Dixon, L., Mueser, K. T., & Torrey, W.
11
12 C. (2001). Implementing evidence-based practices in routine mental health service
13
14 settings. *Psychiatric Services*, 52(2), 179-182.
15
16
- 17 Herdelin, A. C., & Scott, D. L. (1999). Experimental studies of the Program of Assertive
18
19 Community Treatment (PACT): A meta-analysis. *Journal of Disability Policy Studies*, 10,
20
21 53-89.
22
23
- 24 Latimer, E. (1999). Economic impacts of assertive community treatment: A review of the
25
26 literature. *Canadian Journal of Psychiatry*, 44, 443-454.
27
28
- 29 McGrew, J. H., & Bond, G. R. (1995). Critical ingredients of assertive community treatment:
30
31 Judgments of the experts. *Journal of Mental Health Administration*, 22, 113-125.
32
33
- 34 McGrew, J. H., & Bond, G. R. (1997). The association between program characteristics and
35
36 service delivery in assertive community treatment. *Administration and Policy in Mental*
37
38 *Health*, 25, 175-189.
39
- 40 McGrew, J. H., Bond, G. R., Dietzen, L. L., & Salyers, M. P. (1994). Measuring the fidelity of
41
42 implementation of a mental health program model. *Journal of Consulting and Clinical*
43
44 *Psychology*, 62(4), 670-678.
45
46
- 47 McGrew, J. H., Stull, L. G., Rollins, A. L., Salyers, M. P., & Hicks, L. J. (2011). A comparison of
48
49 phone-based and on-site assessment of fidelity for assertive community treatment in
50
51 Indiana. *Psychiatr Serv*, 62(6), 670-674. doi: 10.1176/appi.ps.62.6.670
52
- 53 McGrew, J. H., White, L. M., & Stull, L. G. (2013). Self-assessed fidelity: proceed with caution:
54
55 in reply. *Psychiatr Serv*, 64(4), 394. doi: 10.1176/appi.ps.640419
56
57
58
59
60
61
62
63
64
65

- 1
2
3
4 McGrew, J. H., White, L. M., Stull, L. G., & Wright-Berryman, J. (2013). A comparison of self-
5
6 reported and phone-administered methods of ACT fidelity assessment: a pilot study in
7
8 Indiana. *Psychiatr Serv*, 64(3), 272-276. doi: 10.1176/appi.ps.001252012
9
- 10 McHugo, G. J., Drake, R. E., Teague, G. B., & Xie, H. (1999). Fidelity to assertive community
11
12 treatment and client outcomes in the New Hampshire Dual Disorders Study. *Psychiatric*
13
14 *Services*, 50(6), 818-824.
15
- 16
17 McHugo, G. J., Drake, R. E., Whitley, R., Bond, G. R., Campbell, K., Rapp, C. A., . . . Finnerty,
18
19 M. T. (2007). Fidelity outcomes in the national implementing evidence-based practices
20
21 project. *Psychiatric Services*, 58(10), 1279-1284.
22
23
- 24 Mueser, K. T., Bond, G. R., Drake, R. E., & Resnick, S. G. (1998). Models of community care for
25
26 severe mental illness: A review of research on case management. *Schizophrenia*
27
28 *Bulletin*, 24, 37-74.
29
- 30
31 Neale, M. S., Rosenheck, R., Castrodonatti, J., Martin, A., Morrissey, J., & D'Amico, M. (2007).
32
33 *Mental Health Intensive Case Management (MHICM) in the Department of Veterans*
34
35 *Affairs: The Tenth National Performance Monitoring Report - FY 2006*. West Haven, CT:
36
37 VHA NEPEC Report (203.987.3850).
38
- 39
40 Phillips, S. D., Burns, B. J., Edgar, E. R., Mueser, K. T., Linkins, K. W., Rosenheck, R. A., . . .
41
42 McDonel Herr, E. C. (2001). Moving assertive community treatment into standard
43
44 practice. *Psychiatric Services*, 52(6), 771-779.
45
- 46
47 President's New Freedom Commission on Mental Health. (2003). *Achieving the promise:*
48
49 *Transforming mental health care in America. Final Report. DHHS Pub. No. SMA-03-*
50
51 *3832*. Rockville, MD: Substance Abuse and Mental Health Services Administration.
52
- 53
54 Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the
55
56 efficacy of diagnostic tests. *N Engl J Med*, 299(17), 926-930.
57
- 58
59 Rosenheck, R., Neale, M. S., Leaf, P., Milstein, R., & Frisman, L. (1995). Multisite experimental
60
61 cost study of intensive psychiatric community care. *Schizophrenia Bulletin*, 21, 129-140.
62
63
64
65

- 1
2
3
4 Rosenheck, R. A., & Neale, M. S. (1998). Cost-effectiveness of intensive psychiatric community
5
6 care for high users of inpatient services. *Archives of General Psychiatry*, 55(5), 459-466.
7
8
9 Salyers, M. P., McKasson, R. M., Bond, G. R., McGrew, J. H., Rollins, A. L., & Boyle, C. (2007).
10
11 The Role of Technical Assistance Centers in Implementing Evidence-Based Practices:
12
13 Lessons Learned. *American Journal of Psychiatric Rehabilitation*, 10(2), 85-101.
14
15 Salyers, M. P., Bond, G. R., Teague, G. B., Cox, J. F., Smith, M. E., Hicks, M. L., & Koop, J. I.
16
17 (2003). Is it ACT yet? Real-world examples of evaluating the degree of implementation
18
19 for assertive community treatment. *Journal of Behavioral Health Services and Research*,
20
21 30(3), 304-320.
22
23
24 Stein, L. I., & Test, M. A. (1980). An alternative to mental health treatment. I: Conceptual
25
26 model, treatment program, and clinical evaluation. *Archives of General Psychiatry*, 37,
27
28 392-397.
29
30
31 Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches
32
33 to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4),
34
35 66-78.
36
37
38 Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity in assertive community
39
40 treatment: development and use of a measure. *American Journal of Orthopsychiatry*,
41
42 68(2), 216-232.
43
44
45 Torrey, W. C., Finnerty, M., Evans, A., & Wyzik, P. F. (2003). Strategies for leading the
46
47 implementation of evidence-based practices. *Psychiatric Clinics of North America*, 26,
48
49 883-897.
50
51
52 Ziguras, S., & Stuart, G. (2000). A meta-analysis of the effectiveness of mental health case
53
54 management over 20 years. *Psychiatric Services*, 51, 1410-1421.
55
56
57
58
59
60
61
62
63
64
65