Ensuring Quality of Mental Health Services: Conceptual and Practical Issues of Implementation Fidelity

John McGrew, Ph.D.
Department of Psychology
Indiana University Purdue University Indianapolis

March 27, 2014
Florida State University
Background: The problem

Just because we have a good treatment, doesn’t guarantee that therapists are delivering it or clients are getting it.
The “95% Problem”

- **Limited access to care or no care ➔**
  - 60% without care: mostly dropouts (New Freedom Commission, 2003)

- **Have access, but poor care ➔**
  - 35% with inadequate care: science-to-service gap (Institute of Medicine, 2005)

---

The implementation problem—It’s probably Prozac

An illustrative story: A trip to the drug store

- **Customer** (picking up Prozac): Do you have my Prozac ready?
- **Pharmacist**: Sure, well, it is an enhanced Prozac.
- **Customer**: What do you mean?
- **Pharmacist**: Well, Phil and I have found that if we add some extra ingredients and also shave off a little of some of the “harsher” ingredients it makes a better mix of “Prozac.”
- **Customer**: You mean Prozac bought in one place may not be at all like Prozac bought somewhere else ... but I want the real Prozac, how do I know what you give me will work as well?
- **ANSWER**: TRUST ME!
Fidelity matters! Fidelity and hospital reduction in 18 ACT Teams (McGrew, Bond, Dietzen, Salyers, 1994)

- Percent reduction in hospital use
- Three fidelity scales
  - Total fidelity
  - Staffing fidelity
  - Organizational fidelity
Example program model: Assertive Community Treatment

Hospital without walls
ACT basic elements

- Multidisciplinary staffing
- Team approach
- Integrated services
- Direct service provider (not brokering)
- Low client-staff ratios (10:1)
- More than 75% of contacts in the community
- Assertive outreach
- Focus on symptom management and everyday problems in living
- Ready access in times of crisis
- Time-unlimited services
Outcomes from 25 Experimental Evaluations of ACT (Bond, 2001)

Table 1. Comparison of ACT to Controls in 25 RCTs

<table>
<thead>
<tr>
<th></th>
<th>ACT Compared to Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Better</td>
</tr>
<tr>
<td>Hospital use</td>
<td>17 (74%)</td>
</tr>
<tr>
<td>Housing stability</td>
<td>8 (67%)</td>
</tr>
<tr>
<td>Symptoms</td>
<td>7 (44%)</td>
</tr>
<tr>
<td>Quality of life</td>
<td>7 (58%)</td>
</tr>
</tbody>
</table>

Conceptual issues with fidelity assessment
Fidelity and related concepts

- **Fidelity**—Faithful implementation of an empirically-supported treatment model or adherence to program standards (Bond et al., 2000)
- **Historical precursors** (Moncher & Prinz, 1991)
  - Treatment integrity/treatment adherence
  - Treatment differentiation
- **Experimental validity** (Cook & Campbell, 1991)
  - Construct validity of the independent variable
  - Implementation check
- **Operational definition**
  - Treatment manuals
- **Psychotherapy process research**
  - Critical ingredients
The basic assumption

- Fidelity
- Quality Service
- Mechanisms of action
- Clinical outcomes
Some steps in constructing a fidelity scale

- Identify specific program model
- Identify critical elements of program model
- Identify appropriate (e.g., valid, reliable) sources for measuring elements
- Operationalize elements (i.e., construct measures of critical elements)
- Identify subscales
- Pilot test
- Validation study
OK, we know our program works, but what exactly is working?
Critical ingredients: Some methodological issues

- Models elements usually defined BEFORE empirical testing \(\rightarrow\) pre-scientific (Weston et al., 2004)
- Factors that may impact critical elements
  - Outcome (quality of life, hospital reduction, cost)
  - Setting (urban, rural)
  - Client subgroup (co-morbid substance use)
  - Criterion of criticalness (helpful, essential, unique, critical to an outcome)
  - As judged by whom (experts, clients, clinicians)
- How broadly we cast our net
  - Critical to this EBP only
  - Plus common treatment factors (rapport, empathy)
  - Plus elements critical to quality implementation (organizational culture?)
- How do we determine what is critical?
  - Using what empirical methods (next slide)
Empirical methods to determine critical ingredients

- **Dismantling studies** (vary elements in within study comparisons)
- **Meta-analytic studies** (across study comparisons)
- **Normative standards** (what is implemented most often is more likely to be critical)
- **Stakeholder surveys** (ask experts, consumers)

**NOTE:** Rigor and feasibility of empirical methods tend to be inversely related
ACT Critical ingredients

Example: Meta-analysis

Decreased hospital use

- Shared caseloads \(0.65^{**}\)
- Number of contacts \(0.59^{**}\)
- 24 hour availability \(0.55^*\)
- Daily team meeting \(0.49^*\)
- Nurse on team \(0.49^*\)

Examples: Dismantling

- Single case manager vs. Team approach
  - Team approach leads to more stable hospital reductions (Bond, Pensec et al., 1991)
- Low vs Hi Caseload ratios
  - Lower caseloads \(\rightarrow\) better outcomes (Jerrell, 1999)
- Peer counselors vs. non-peer counselors
  - Mixed results


Implementation vs. Intervention fidelity

Inside the Black Box: a model of ACT helping

Implementation

- Organizational ingredients
- Structural ingredients

Intervention

- Clinician actions/intervention
- Medication management
- Helping Alliance
- Social network support

Mechanisms of action
### ACT workers’ perspectives on clinical ingredients: Top ten ingredients
(N=73; McGrew et al., 2003)

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication management</td>
<td>1.19</td>
</tr>
<tr>
<td>Continuing assessment</td>
<td>1.38</td>
</tr>
<tr>
<td>Regular home visits</td>
<td>1.45</td>
</tr>
<tr>
<td>Problem-solving support</td>
<td>1.52</td>
</tr>
<tr>
<td>Shared caseloads</td>
<td>1.55</td>
</tr>
<tr>
<td>Access to medical care</td>
<td>1.66</td>
</tr>
<tr>
<td>Adequate housing</td>
<td>1.73</td>
</tr>
<tr>
<td>Provision of social support</td>
<td>1.87</td>
</tr>
<tr>
<td>Money management</td>
<td>2.00</td>
</tr>
<tr>
<td>Increase in social contacts</td>
<td>2.05</td>
</tr>
</tbody>
</table>

1=very beneficial, 7=not at all beneficial
Practical issues with fidelity assessment
Fidelity harder to achieve for some EBPs: National EBP Project 2-Year Rates of Successful Program Implementation

<table>
<thead>
<tr>
<th></th>
<th>Successful (Fidelity &gt;4)</th>
<th>Unsuccessful</th>
<th>Dropped Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>10 (77%)</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>8 (89%)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>IDDT</td>
<td>2 (15%)</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>IMR</td>
<td>6 (50%)</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>FPE</td>
<td>3 (50%)</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>29 (55%)</td>
<td>20</td>
<td>4</td>
</tr>
</tbody>
</table>

- EBPs differed in:
  - Clinical complexity
  - Practitioner familiarity
  - Compatibility with usual practice
## Key difference: Type of fidelity items

<table>
<thead>
<tr>
<th>Structural Fidelity Items</th>
<th>Assessing clinical interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Things that can be done by administrative fiat, such as:</td>
<td>- Practitioner actions that follow prescribed techniques, such as:</td>
</tr>
<tr>
<td>- Daily team meetings</td>
<td>- Motivational interviewing</td>
</tr>
<tr>
<td>- Multidisciplinary staffing</td>
<td>- Behavioral tailoring</td>
</tr>
<tr>
<td>- Low caseload ratio</td>
<td>- Providing stagewise interventions</td>
</tr>
<tr>
<td>- Following a curriculum</td>
<td></td>
</tr>
<tr>
<td>- Distributing educational handouts</td>
<td></td>
</tr>
</tbody>
</table>
Fidelity Burden—The elephant in the room: Explosion of interest in EBPs
Current models for fidelity assessment are very time intensive

- It is nearly universally accepted that EBPs require fidelity monitoring to ensure accurate implementation.
- The gold standard for implementation fidelity monitoring is onsite (or reviewing of tapes for intervention fidelity) which requires considerable assessment time for both assessor and agency (as much as 2-3 days).
- The burden to the credentialing body, usually the state authority, increases exponentially with:
  - The number of potential EBPs.
  - The number of sites adopting each EBP.
There are too many EBPs for current models of fidelity monitoring

<table>
<thead>
<tr>
<th>Date</th>
<th>Review source</th>
<th>Number of EBPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>Division 12 Taskforce</td>
<td>22 effective, 7 probable</td>
</tr>
<tr>
<td>1998</td>
<td>Treatments that Work</td>
<td>44 effective, 20 probable</td>
</tr>
<tr>
<td>2001</td>
<td>National EBP Project</td>
<td>6 effective</td>
</tr>
<tr>
<td>2001</td>
<td>Chambless, Annual Review of Psychology Article</td>
<td>108 effective or probable for adults; 37 for children</td>
</tr>
<tr>
<td>2005</td>
<td>What works for whom</td>
<td>31 effective, 28 probable</td>
</tr>
<tr>
<td>2007</td>
<td>Treatments that Work</td>
<td>69 effective, 73 probable</td>
</tr>
<tr>
<td>2014</td>
<td>Division 12, APA</td>
<td>79 effective</td>
</tr>
<tr>
<td>2014</td>
<td>SAMHSA Registry</td>
<td>88 experimental, replicated programs</td>
</tr>
</tbody>
</table>
Alternative quality assurance mechanisms to alleviate the assessment burden*

- Use of shorter scales (NOTE: both the newly revised DACTS and IPS scales are longer)
- Increase length of time between fidelity assessments
- Use of need-based vs. fixed interval schedules of assessment
- Use of alternative methods of assessment (e.g., self report, phone)

## Factors impacting fidelity assessment

<table>
<thead>
<tr>
<th>Mode of collection</th>
<th>Face-to-face, Phone, Self-report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designated rater</td>
<td>Independent rater, provider</td>
</tr>
<tr>
<td>Data collection site</td>
<td>On-site</td>
</tr>
<tr>
<td></td>
<td>Off-site</td>
</tr>
<tr>
<td>Data collector</td>
<td>External—outside assessor</td>
</tr>
<tr>
<td></td>
<td>Agency affiliated—within agency, but outside the team</td>
</tr>
<tr>
<td></td>
<td>Internal—self assessment by team/program</td>
</tr>
<tr>
<td>Instrument</td>
<td>Full/ partial/ screen</td>
</tr>
<tr>
<td>Data source</td>
<td>EMR, chart review, self-report, observation</td>
</tr>
<tr>
<td>Informants</td>
<td>Team leader, full team, specific specialties (e.g., nurse), clients, significant others</td>
</tr>
<tr>
<td>Site variables potentially impacting</td>
<td>Size, location, years of operation, developmental status</td>
</tr>
</tbody>
</table>
Reducing burden: Fidelity assessment for Assertive Community Treatment
“Gold standard” fidelity scale for ACT: Dartmouth Assertive Community Treatment Scale (DACTS)

- 28-item scale, 5-point behaviorally-anchored scale (1=not implemented to 5=full implementation)
- Three subscales:
  - **Human Resources Subscale** (11 items) Small caseload, team approach, psychiatrist, nurse
  - **Organizational Boundaries Subscale** (7 items) Admission criteria, hospital admission/discharge, crisis services
  - **Nature of Services Subscale** (10 items) Community-based services, no dropout policy, intensity of services, frequency of contact

DACTS Scoring

- **Individual Items**
  - Rating of $\leq 3$ = Unacceptable implementation
  - Rating of $4$ = Acceptable/good implementation
  - Rating of $5$ = Excellent implementation

- **Subscale scores and Total score**
  - Mean of $\leq 4.0$ = Below acceptable standards for adherence to model
  - Mean of $4.0 - 4.3$ = Good adherence to model
  - Mean of $\geq 4.3$ = Exemplary adherence to model
<table>
<thead>
<tr>
<th>DACTS Items</th>
<th>Ancehrs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Resources Items</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>H1 SMALL CASELOAD: client/provider ratio of 10:1.</td>
<td>50 clients/clinician or more.</td>
</tr>
<tr>
<td>H2 TEAM APPROACH: Provider group functions as team rather than as individual practitioners; clinicians know and work with all clients.</td>
<td>Fewer than 10% clients with multiple staff face-to-face contacts in 2-weeks</td>
</tr>
<tr>
<td>H3 PROGRAM MEETING: Program meets frequently to plan and review services for each client.</td>
<td>Program service-planning for each client usually occurs once/month or less frequently</td>
</tr>
<tr>
<td>H4 PRACTICING TEAM LEADER: Supervisor of front line clinicians provides direct services.</td>
<td>Supervisor provides no services.</td>
</tr>
<tr>
<td>H5 CONTINUITY OF STAFFING: program maintains same staffing over time.</td>
<td>Greater than 80% turnover in 2 years.</td>
</tr>
<tr>
<td>H6 STAFF CAPACITY: Program operates at full staffing.</td>
<td>Program has operated at less than 50% of staffing in past 12 months.</td>
</tr>
<tr>
<td>H7 PSYCHIATRIST ON STAFF: there is at least one full-time psychiatrist per 100 clients assigned to work with the program.</td>
<td>Program for 100 clients has less than .10 FTE regular psychiatrist.</td>
</tr>
</tbody>
</table>
Study 1: Phone Based Assessment
Why phone based? Preliminary studies demonstrating predictive validity

<table>
<thead>
<tr>
<th></th>
<th>Correlations between closure rates and total fidelity scores inSupported Employment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QSEIS and VR closure rates</td>
</tr>
<tr>
<td>McGrew &amp; Griss, 2005, n=23</td>
<td>.42*</td>
</tr>
<tr>
<td>McGrew, 2007, n=17</td>
<td>n/a</td>
</tr>
<tr>
<td>McGrew, 2008, n=23</td>
<td>n/a</td>
</tr>
</tbody>
</table>
A comparison of phone-based and onsite-based fidelity for ACT: Research questions

- Compared to onsite, is phone based fidelity assessment
  - Reliable
  - Valid
  - With reduced burden
- Does rater expertness or prior site experience influence fidelity reliability or validity?

A comparison of phone-based and onsite-based fidelity for ACT: Methods

- Design: Within site comparison
- Target sample: 30 ACT teams in Indiana
- Timeframe: One-year accrual
- Phase 1: Develop Phone Protocol
- Phase 2: Test Phone Based vs. Onsite DACTS
  - Completed within one month prior to scheduled onsite
  - For half of the sites: experienced rater plus inexperienced rater
  - For other half: experienced rater plus onsite trainer
  - Interview limited to Team Leader
Development of phone protocol

- Assumptions
  - People tell the truth
  - People want to look good

- Construction guidelines
  - The more molecular, concrete or objective the data, the lower the likelihood of measurement error
  - The more global, interpretive or subjective the data, the greater the likelihood of measurement error
FORMAT USING SUBJECTIVE ESTIMATES

What percent of hospital admissions involve the team?

What percent of the time is the team involved in hospital discharge planning?

<table>
<thead>
<tr>
<th>Client</th>
<th>Admission – team involved?</th>
<th>Discharge – team involved?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>Team brought client into ER and helped with inpatient admission documentation</td>
<td>Team participated in discharge planning prior to release, transported him home upon release</td>
</tr>
<tr>
<td>Client 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Format using subjective estimates

Which of the following services does your program have full responsibility for and provide directly: psychiatric services, counseling/psychotherapy, housing support, substance abuse treatment, employment/rehabilitative services?

Phone interview format

**Table 6. Services Received Outside of ACT Team**
Now review your entire caseload and provide a rough estimate of the number of individuals who have received assistance in the following areas from non-ACT team personnel or providers **during the past 4 weeks**.

<table>
<thead>
<tr>
<th>Service</th>
<th>Number of clients that receive the following services from outside the ACT team (e.g., from residential program, from other program in agency, from program outside agency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living in supervised living situation</td>
<td></td>
</tr>
<tr>
<td>Other housing support outside the ACT team</td>
<td></td>
</tr>
<tr>
<td>Psychiatric services</td>
<td></td>
</tr>
<tr>
<td>Case management</td>
<td></td>
</tr>
<tr>
<td>Counseling/ individual supportive therapy</td>
<td></td>
</tr>
<tr>
<td>Substance abuse treatment</td>
<td></td>
</tr>
<tr>
<td>Employment services</td>
<td></td>
</tr>
<tr>
<td>Other rehabilitative services</td>
<td></td>
</tr>
</tbody>
</table>
Procedure: Phone Fidelity

- Phone interviews via conference call between two raters and TLs
  - Reviewed tables for accuracy
  - Asked supplemental questions
  - Filled in any missing data from self-report protocol

- Initial scoring
  - Raters independently scored the DACTS based on all available information

- Consensus scoring
  - Discrepant items identified
  - Raters met to discuss and reach final consensus scores
Phase 1—Table construction: Results

- Piloted with two VA MHICM teams
- Final Phone protocol includes 9 tables
  - Staffing
  - Client discharges (past 12 months)
  - Client admissions (past 6 months)
  - Recent hospitalizations (last 10)
  - Case review from charts (10 clients) or EMR (total caseload)(frequency/intensity)
  - Services received outside ACT team
  - Engagement mechanisms
  - Miscellaneous (program meeting, practicing TL, crisis, informal supports)
  - IDDT items
Phase 2 Phone based assessment is reliable—interrater reliability

<table>
<thead>
<tr>
<th>Comparison – total DACTS scores</th>
<th>Single Measure ICC</th>
<th>Average Measure ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced rater vs. second rater</td>
<td>0.91</td>
<td>0.93</td>
</tr>
<tr>
<td>ONSITE published estimate*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparing consultant, trainer and implementation monitor</td>
<td>0.99&lt;sup&gt;1&lt;/sup&gt;</td>
<td></td>
</tr>
</tbody>
</table>


Note 1. Type of ICC not specified
Results: Phone based assessment is valid compared to onsite (consistency)

<table>
<thead>
<tr>
<th>Comparisons using DACTS Total Score</th>
<th>Single Measures ICC</th>
<th>Average Measures ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onsite vs. Phone Consensus</td>
<td>0.87</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Phone based had adequate validity compared to onsite for total and subscale scores (consensus)

<table>
<thead>
<tr>
<th>Item/Subscale</th>
<th>Phone Consensus Mean/SD (n = 17)</th>
<th>Onsite Mean/SD (n = 17)</th>
<th>Mean Absolute Difference (n = 17)</th>
<th>Range of Absolute Differences</th>
<th>Intraclass Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total DACTS</td>
<td>4.29 (0.19)</td>
<td>4.30 (0.13)</td>
<td>0.07</td>
<td>0.00 – 0.32</td>
<td>0.87</td>
</tr>
<tr>
<td>Organizational Boundaries</td>
<td>4.72 (0.19)</td>
<td>4.74 (0.18)</td>
<td>0.08</td>
<td>0.00 – 0.29</td>
<td>0.73</td>
</tr>
<tr>
<td>Human Resources</td>
<td>4.35 (0.22)</td>
<td>4.34 (0.28)</td>
<td>0.12</td>
<td>0.00 – 0.27</td>
<td>0.87</td>
</tr>
<tr>
<td>Services</td>
<td>3.91 (0.31)</td>
<td>3.95 (0.23)</td>
<td>0.14</td>
<td>0.00 – 0.50</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Frequency distribution of differences between onsite and phone total DACTS scores
## DACTS Phone Assessment Burden

<table>
<thead>
<tr>
<th>Task</th>
<th>Time (Mean/SD)</th>
<th>Time Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Preparation for call</td>
<td>7.5 hours (6.2)</td>
<td>1.75 to 25</td>
</tr>
<tr>
<td>Phone call</td>
<td>72.8 minutes (18.5)</td>
<td>40 to 111</td>
</tr>
</tbody>
</table>
Explaining the results: Reliability tends to improve over time

<table>
<thead>
<tr>
<th>Comparisons using DACTS Total Score</th>
<th>Single Measures ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced vs. Second rater (1st 8 sites)</td>
<td>0.88</td>
</tr>
<tr>
<td>Experienced vs. Second rater (Last 9 sites)</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Explaining the differences: Rater expertness or prior experience with the site does not influence interrater *reliability*

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Experienced Phone M/SD</th>
<th>Comparison Rater Phone M/SD</th>
<th>Mean Absolute Difference</th>
<th>Range of Absolute Differences</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced vs. Rater 2</td>
<td>4.29 (0.18)</td>
<td>4.31 (0.19)</td>
<td>0.06</td>
<td>0.00 – 0.25</td>
<td>0.91</td>
</tr>
<tr>
<td>Experienced vs. Trainer</td>
<td>4.38 (0.14)</td>
<td>4.44 (0.14)</td>
<td>0.08</td>
<td>0.00 – 0.25</td>
<td>0.92</td>
</tr>
<tr>
<td>Experienced vs. Naïve</td>
<td>4.21 (0.19)</td>
<td>4.19 (0.16)</td>
<td>0.05</td>
<td>0.00 – 0.14</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Explaining the differences: 
Rater prior experience/expertness may influence concurrent *validity* (consistency, but not consensus)

<table>
<thead>
<tr>
<th>Rater</th>
<th>Phone Means/SD</th>
<th>Onsite Means/SD</th>
<th>Mean Absolute Difference (n = 17)</th>
<th>Range of Absolute Differences</th>
<th>Intraclass Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trainer (n=8)</td>
<td>4.44 (0.94)</td>
<td>4.40 (0.95)</td>
<td>0.06</td>
<td>0.00 – 0.32</td>
<td>0.92</td>
</tr>
<tr>
<td>Experienced (n=17)</td>
<td>4.29 (1.03)</td>
<td>4.30 (1.01)</td>
<td>0.07</td>
<td>0.00 – 0.25</td>
<td>0.86</td>
</tr>
<tr>
<td>Inexperienced (n=9)</td>
<td>4.19 (1.06)</td>
<td>4.25 (1.05)</td>
<td>0.08</td>
<td>0.00 – 0.29</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Qualitative results

- Self-report data mostly accurate
- Teams prefer table format
- Teams' concerns/suggestions
  - Phone may limit contact with trainers (limits training opportunities & ecological validity of assessment)
  - Suggestion to involve other members of team, especially substance abuse specialist
Conclusions

- Objective, concrete assessment tends to lead to reliable and valid phone fidelity
  - Most programs classified within .10 scale points of onsite total DACTS
  - Error differences show little evidence of systematic bias (over- or under-estimates)
- Few changes made from self-report tables → objective self-report may account for most of findings
- Raters/rating experience may influence reliability and validity of data collected
  - Ongoing training and rating calibration likely critical
- Large reduction in burden for assessor, modest reduction for site, with a small and likely acceptable degradation in validity
Study 2: Self-report fidelity
Self-report vs Phone Fidelity Study

- **Research question**: Is self-report a useful and less burdensome alternative fidelity assessment method
- **Design**: Compare phone-based fidelity to self-report fidelity
- **Inclusion Criteria**: ACT teams contracted with Indiana Division of Mental Health and Addiction
  - 16 (66.7%) teams agreed; 8 (33.3%) declined to participate

Procedure

- **Phone Fidelity**: same as prior study
- **Self-Report Fidelity**: Two additional raters scored DACTS using information from Self-report Protocol
  - Ratings conducted after completion of all phone interviews
  - Raters not involved in phone interviews and did not have access to information derived from interviews
  - Exception: Two cases where missing data provided before the phone call
- Same scoring procedure as phone fidelity, except scoring based solely on information from self-report protocol
Preliminary results

- Phone interviews averaged 51.4 minutes (SD = 13.6)
- Ranged from 32 to 87 minutes
- Missing data for 9 of 16 (56.3%) teams
  - Phone
    - Raters were able to gather missing data
  - Self-report
    - Raters left DACTS items blank (unscored) if information was missing or unclear
Phone fidelity reliability is excellent (consistency and consensus)

<table>
<thead>
<tr>
<th>Reliability comparisons (n=16)</th>
<th>Experienced Rater</th>
<th>Naïve Rater</th>
<th>Mean Absolute Difference</th>
<th>Range of Absolute Differences</th>
<th>Intraclass Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Total DACTS (Experienced vs. Second Rater)</td>
<td>4.22</td>
<td>.25</td>
<td>4.20</td>
<td>.28</td>
<td>.04</td>
</tr>
<tr>
<td>Organizational Bound. Subscale</td>
<td>4.58</td>
<td>.14</td>
<td>4.57</td>
<td>.14</td>
<td>.06</td>
</tr>
<tr>
<td>Human Resources Subscale</td>
<td>4.27</td>
<td>.35</td>
<td>4.30</td>
<td>.36</td>
<td>.05</td>
</tr>
<tr>
<td>Nature of Services Subscale</td>
<td>3.91</td>
<td>.41</td>
<td>3.84</td>
<td>.46</td>
<td>.07</td>
</tr>
</tbody>
</table>

Differences of ≤ .25 (5% of scoring protocol)

- Total DACTS: Differences < .25 for all 16 sites
- Organizational Boundaries: Differences < .25 for 16 sites
- Human Resources: Differences < .25 for 15 of 16 sites
- Nature of Services: Differences < .25 for 15 of 16 sites
Self-report fidelity reliability ranges from good to poor

<table>
<thead>
<tr>
<th>Reliability comparisons (n=16)</th>
<th>Consultant Rater</th>
<th>Experienced Rater</th>
<th>Mean Absolute Difference</th>
<th>Range of Absolute Differences</th>
<th>Intraclass Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Total DACTS</td>
<td>4.16</td>
<td>.27</td>
<td>4.11</td>
<td>.26</td>
<td>.14</td>
</tr>
<tr>
<td>Organizational Bound. Subscale</td>
<td>4.49</td>
<td>.20</td>
<td>4.53</td>
<td>.21</td>
<td>.13</td>
</tr>
<tr>
<td>Human Resources Subscale</td>
<td>4.27</td>
<td>.39</td>
<td>4.21</td>
<td>.28</td>
<td>.25</td>
</tr>
<tr>
<td>Nature of Services Subscale</td>
<td>3.72</td>
<td>.50</td>
<td>3.76</td>
<td>.48</td>
<td>.20</td>
</tr>
</tbody>
</table>

Absolute differences between raters (consensus) were moderate

- Total DACTS: Differences < .25 for 13 sites
- Organizational Boundaries: Differences < .25 for 13 sites
- Human Resources: Differences < .25 for 10 sites
- Nature of Services: Differences < .25 for 11 sites
Validity of self-report vs phone fidelity is good to acceptable (consistency and consensus)

<table>
<thead>
<tr>
<th>Validity comparisons (n=16)</th>
<th>Self-Report</th>
<th>Phone</th>
<th>Mean Absolute Difference</th>
<th>Range of Absolute Differences</th>
<th>Intraclass Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Total DACTS</td>
<td>4.12</td>
<td>.27</td>
<td>4.21</td>
<td>.27</td>
<td>.13</td>
</tr>
<tr>
<td>Organizational Bound. Subscale</td>
<td>4.53</td>
<td>.15</td>
<td>4.56</td>
<td>.12</td>
<td>.08</td>
</tr>
<tr>
<td>Human Resources Subscale</td>
<td>4.22</td>
<td>.31</td>
<td>4.29</td>
<td>.34</td>
<td>.15</td>
</tr>
<tr>
<td>Nature of Services Subscale</td>
<td>3.72</td>
<td>.49</td>
<td>3.87</td>
<td>.47</td>
<td>.20</td>
</tr>
</tbody>
</table>

Absolute differences between methods (consensus) were small to medium
- Total DACTS: Differences < .25 for 15 or 16 sites
- Organizational Boundaries: Differences < .25 for 15 sites
- Human Resources: Differences < .25 for 10 sites
- Nature of Services: Differences < .25 for 12 sites
## Problematic Items

**Mean absolute differences of .25 or higher (5% of scoring range)**

<table>
<thead>
<tr>
<th>Items</th>
<th>Subscale</th>
<th>Self-Report</th>
<th>Phone</th>
<th>Difference</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual Diagnosis Model</td>
<td>Nature of Services</td>
<td>3.80</td>
<td>4.56</td>
<td>.76</td>
<td>t = 4.58, p &lt; .001</td>
</tr>
<tr>
<td>Vocational Specialist</td>
<td>Human Resources</td>
<td>3.25</td>
<td>3.88</td>
<td>.63</td>
<td>t = 1.67, p = .116</td>
</tr>
<tr>
<td>Informal Support System</td>
<td>Nature of Services</td>
<td>3.00</td>
<td>3.44</td>
<td>.44</td>
<td>t = 1.60, p = .130</td>
</tr>
<tr>
<td>Responsibility for Crisis Services</td>
<td>Organizational Boundaries</td>
<td>4.31</td>
<td>4.69</td>
<td>.38</td>
<td>t = 3.00, p = .009</td>
</tr>
<tr>
<td>Consumer on Team</td>
<td>Nature of Services</td>
<td>1.75</td>
<td>1.38</td>
<td>.37</td>
<td>t = -1.38, p = .189</td>
</tr>
<tr>
<td>Responsibility for Tx Services</td>
<td>Organizational Boundaries</td>
<td>4.44</td>
<td>4.69</td>
<td>.25</td>
<td>t = 2.23, p = .041</td>
</tr>
<tr>
<td>Continuity of Staff</td>
<td>Human Resources</td>
<td>3.31</td>
<td>3.06</td>
<td>.25</td>
<td>t = 1.379, p = .188</td>
</tr>
</tbody>
</table>
Classification: Sensitivity and Specificity

ACT Team = Fidelity Score ≥ 4.0, Phone=criterion

<table>
<thead>
<tr>
<th></th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACT Team</td>
</tr>
<tr>
<td>Self-Report</td>
<td></td>
</tr>
<tr>
<td>ACT Team</td>
<td>10</td>
</tr>
<tr>
<td>Not ACT Team</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
</tr>
</tbody>
</table>

Sensitivity = .77  
Specificity = 1.00  
Predictive Power = .81  
False Positive Rate = .00  
False Negative Rate = .23
Preliminary conclusions

- Support for reliability and validity of self-report fidelity, especially for total score
- Self-report assessment in agreement (\(\leq .25\) scale points) with phone assessment for 94% of sites

- Self-report fidelity assessment viable for gross, dichotomous judgments of adherence
- No evidence of inflated self reporting
  - Self-report fidelity underestimated phone fidelity for 12 (75%) sites
Study 3: Preliminary results—Comparison of four methods of fidelity assessment (n=32)

- 32 VA MHICM sites
- Contrasted four fidelity methods
  - Onsite
  - Phone
  - Self-report—objective scoring
  - Self-assessment
- Addresses concerns from prior studies:
  - sampling limited to fidelity experienced, highly adherent teams in single state
  - failure to use onsite as comparison criterion
Validity of phone vs onsite fidelity good

<table>
<thead>
<tr>
<th>Validity comparisons (n=32)</th>
<th>Onsite</th>
<th></th>
<th>Phone</th>
<th></th>
<th>Mean Absolute Difference</th>
<th>Range of Absolute Differences</th>
<th>Intraclass Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total DACTS</td>
<td>3.22</td>
<td>.28</td>
<td>3.15</td>
<td>.28</td>
<td>.13</td>
<td>.00 – 0.50</td>
<td>.88</td>
</tr>
<tr>
<td>Organizational Bound. Subscale</td>
<td>3.76</td>
<td>.38</td>
<td>3.64</td>
<td>.35</td>
<td>.18</td>
<td>.00 – 0.80</td>
<td>.85</td>
</tr>
<tr>
<td>Human Resources Subscale</td>
<td>3.38</td>
<td>.41</td>
<td>3.35</td>
<td>.43</td>
<td>.16</td>
<td>.00 – 0.70</td>
<td>.94</td>
</tr>
<tr>
<td>Nature of Services Subscale</td>
<td>2.66</td>
<td>.33</td>
<td>2.60</td>
<td>.31</td>
<td>.18</td>
<td>.00 – 0.70</td>
<td>.84</td>
</tr>
</tbody>
</table>
Validity of self-report vs. onsite is good to acceptable

<table>
<thead>
<tr>
<th>Validity comparisons (n=32)</th>
<th>Onsite</th>
<th>Self-report</th>
<th>Mean Absolute Difference</th>
<th>Range of Absolute Differences</th>
<th>Intraclass Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Total DACTS</td>
<td>3.22</td>
<td>.28</td>
<td>3.17</td>
<td>.31</td>
<td>.17</td>
</tr>
<tr>
<td>Organizational Bound. Subscale</td>
<td>3.76</td>
<td>.38</td>
<td>3.62</td>
<td>.40</td>
<td>.26</td>
</tr>
<tr>
<td>Human Resources Subscale</td>
<td>3.38</td>
<td>.41</td>
<td>3.35</td>
<td>.48</td>
<td>.19</td>
</tr>
<tr>
<td>Nature of Services Subscale</td>
<td>2.66</td>
<td>.33</td>
<td>2.66</td>
<td>.40</td>
<td>.25</td>
</tr>
</tbody>
</table>
General conclusions

• Phone fidelity
  • Good reliability and good to acceptable validity
  • Burden is much less for assessor and reduced for provider

• Self-report fidelity
  • Adequate to fair reliability and good to fair validity
  • More vulnerable to missing data
  • Burden reduced for both assessor and provider vs. phone

• But, support for alternate methods is controversial

Some additional concerns with fidelity measurement

- **External Validity**: Generalizability for different samples and across time (new vs. established teams)
- **Construct Validity**: Are items eminence based or evidence based?
  - TMACT vs DACTS
  - SE Fidelity Scale vs. IPS scale

Implications for Future

- Onsite is impractical as sole or primary method
- All three methods can be integrated into a hierarchical fidelity assessment approach
  - Onsite fidelity for assessing new teams or teams experiencing a major transition
  - Phone or self-report fidelity for monitoring stable, existing teams

Fidelity Assessment System

New Program?

YES
Onsite Visit

NO
Self Report below 4.0
Phone Interview
Score below 4.0
Onsite Visit
Score above 4.0
Phone Interview

Self Report Above 4.0
Alarm Bells?
YES
Phone Interview
NO
Self Report
Big picture: Fidelity is only part of larger set of strategies for assessing and ensuring quality

- Policy and administration
  - Program standards
  - Licensing & certification
  - Financing
  - Dedicated leadership
- Training and consultation
  - Practice-based training
  - Ongoing consultation
  - Technical assistance centers
- Operations
  - Selection and retention of qualified workforce
  - Oversight & supervision
  - Supportive organizational climate /culture
- Program evaluation
  - Outcome monitoring
  - Service-date monitoring
  - Fidelity assessment

Monroe-Devita et al. (2012). Program fidelity and beyond: Multiple strategies and criteria for ensuring quality of Assertive Community Treatment. Psychiatric Services, 63, 743-750.
An alternate to fidelity

- Skip the middleman
- Measure outcomes directly
  - Pay for performance
  - Outcome feedback/management
  - Benchmarking
- Report cards


Results Based Funding: Milestone Attainment Across Sites

**p < .05, **p < .01
Performance tracking
Alternate to fidelity: Outcome management

Thanks to the following collaborators!

- Angie Rollins
- Michelle Salyers
- Alan McGuire
- Lia Hicks
- Hea-Won Kim
- David McClow
- Jennifer Wright-Berryman
- Laura Stull
- Laura White
Thanks for your attention!
IUPUI and Indianapolis: Stop by and visit!
EXTRA SLIDES
Welcome to Indianapolis!
That’s all for now!

Questions??
Explaining the differences: Are errors smaller for high fidelity items?

<table>
<thead>
<tr>
<th></th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Resources Subscale</td>
<td>-0.83**</td>
</tr>
<tr>
<td>Organizational Boundaries Subscale</td>
<td>-0.67**</td>
</tr>
<tr>
<td>Services Subscale</td>
<td>-0.58* (0.27)¹</td>
</tr>
<tr>
<td>Total DACTS</td>
<td>-0.74** (-0.34)¹</td>
</tr>
</tbody>
</table>

* p<.05; ** p<.01

Time difference: range = 1 – 22 days; M(SD) = 5.61(5.49)

Note 1: includes S10–peer specialist
## Phone Fidelity

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong Reliability</td>
<td>Time intensive</td>
</tr>
<tr>
<td>Strong validity with onsite visit(^\text{16})</td>
<td>Scheduling issues</td>
</tr>
<tr>
<td>Less burdensome than onsite visit</td>
<td>Less comprehensive than onsite fidelity visit</td>
</tr>
<tr>
<td>Gathers more detailed information than self-report</td>
<td>May be redundant with self-report fidelity</td>
</tr>
<tr>
<td>Identifies missing data</td>
<td></td>
</tr>
<tr>
<td>Personal communication with TL (and other members of team)</td>
<td></td>
</tr>
<tr>
<td>Opportunity to discuss issues, problems, feedback, etc.</td>
<td></td>
</tr>
</tbody>
</table>
Self-Report Fidelity

**Strengths**
- Least burdensome form of fidelity assessment
- Time efficient
- Acceptable validity with phone fidelity
- Good classification accuracy
- Ensures review and discussion of services among team members
- Explicit protocol to serve as guideline for teams

**Weaknesses**
- Moderate reliability
- Missing Data
- Underestimates true level of fidelity
- Less detailed information than phone or onsite visit
- Not sensitive to item-level problems
- No opportunity to discuss services, issues, feedback with raters
Figure 1. Treatment and service evaluation. CPG, clinical practice guidelines; RCT, randomized controlled trial; TQM, total quality management.
FIGURE 1. Research pipeline.

The 17-year odyssey

Priorities for research funding

Peer review of grants

Publication priorities and peer review

Research synthesis

Guidelines for evidence-based practice

Evidence-based medicine movement

Academic appointments, promotion, and tenure criteria

Funding; population needs, demands; local practice circumstances; professional discretion; credibility and fit of the evidence.
Program Quality Assurance Model
(Feedback Loop)

Participants
Program Leaders
District Administration
Principle Investigator / Directors
External Evaluator

Shewhart, 1939
Abbreviated Measures
Alternate Fidelity Methods: Shorter scales

- Shorter scales take less time to administer
- Short scales have a variety of potential uses:
  - Screens
  - Estimates of full scale
  - Signal/trigger indicators
- Key issue: Selected items may work differently within different samples or at different times:
  - Discriminate ACT from non-ACT in mixed sample of case management programs
  - Discriminate level of ACT fidelity in sample of mostly ACT teams
  - Discriminate in new teams vs. established teams
Identification of DACTS Items for abbreviated scale: Methods

- Four samples used:
  - Salyers et al. (2003), n=87, compares ACT, ICM and BRK
  - Winters & Calsyn (2000), n=18, ACCESS study homeless teams
  - McGrew (2001), n=35, 16-State Performance Indicators, mixed CM teams
  - ACT Center (2001-2008), n=32, ACT teams at 0, 6, 12, 18 and 24 months

- Two criterion indicators:
  - ability to discriminate between known groups
  - correlation to total DACTS
| H1 | Small caseload | 29.6 | 0.62 | 0.46 | 3 |
| H2 | Team approach  | 14.9 | 0.55 | 0.32 | 2 |
| H3 | Program meeting|        |      |      | 0 |
| H4 | Practicing Leader| 0.43 |      | 0.64 | 4 |
| H5 | Staff Continuity|        |      |      | 0 |
| H6 | Staff Capacity  |        |      |      | 0 |
| H7 | Psychiatrist    |        | 0.62 | 0.5  | 2 |
| H8 | Nurse           | 14.2  | 0.72 | 0.41 | 3 |
| H9 | SA Specialist   | 14.2  | 0.56 |      | 1 |
| H10| Voc Specialist  |        | 0.5  |      | 1 |
| H11| Program size    | na    |      | 0.62 | 1 |
| O1 | Admission criteria| 39.4 | 0.36 | 0.66 | 3 |
| O2 | Intake rate     | 18.2  |      |      | 1 |
| O3 | Full responsibility| 25.5 | 0.45 | 0.49 | 4 |
| O4 | Crisis services |        | 0.65 |      | 1 |
| O5 | Involved in hosp admits| |      | 0.38 | 1 |
| O6 | Involved in hosp dischg| | 0.39 |      | 1 |
| O7 | Graduation rate | 15.4  |      |      | 1 |
| S1 | In vivo services| 12.9  |      |      | 1 |
| S2 | Dropouts        |        |      |      | 0 |
| S3 | Engagement mech |        | 0.46 |      | 1 |
| S4 | Service intensity| 18.3 | 0.43 | 0.48 | 3 |
| S5 | Contact frequency|      | 0.38 | 0.54 | 3 |
| S6 | Informal supports| 15.1 | 0.39 | 0.33 | 3 |
| S7 | Indiv SA Tx     | 0.36  |      |      | 1 |
| S8 | DD groups       |        |      |      | 0 |
| S9 | DD model        |        | 0.4  |      | 1 |
| S10| Peer specialists| na    |      |      | 0 |
Abbreviated DACTS Items

- Seven items in “top 10” across 4 different samples
  - Small caseloads (H1)
  - Nurse on team (H8)
  - Clear, consistent, appropriate admission criteria (O1)
  - Team takes full responsibility for services (O3)
  - High service intensity (hours) (S4)
  - High service frequency (contacts) (S5)
  - Frequent contact with informal supports (S6)
**DACTS screen vs. DACTS (cut score = 4)**

<table>
<thead>
<tr>
<th>DACTS screen</th>
<th>16 State</th>
<th>ACT Center</th>
<th>ACT Center</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Baseline</td>
<td>Follow-up</td>
</tr>
<tr>
<td>ACT</td>
<td>7</td>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>Non-ACT</td>
<td>3</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Non-ACT</td>
<td>1</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>14</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation with DACTS</th>
<th>.86</th>
<th>.86</th>
<th>.83</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>.88</td>
<td>1.0</td>
<td>.91</td>
</tr>
<tr>
<td>Specificity</td>
<td>.89</td>
<td>.64</td>
<td>.71</td>
</tr>
<tr>
<td>PPP</td>
<td>.70</td>
<td>.53</td>
<td>.92</td>
</tr>
<tr>
<td>NPP</td>
<td>.96</td>
<td>1.0</td>
<td>.68</td>
</tr>
<tr>
<td>Overall PP</td>
<td>.89</td>
<td>.74</td>
<td>.87</td>
</tr>
</tbody>
</table>

Sensitivity=True Positives; Specificity=True Negatives; PPP = % correct screened positive; NPP = % correct screened negative; OPP=correct judgments/total
Abbreviated DACTS summary

- Findings very preliminary
- Stable, high correlation with overall DACTS
- Overall predictive power acceptable to good (.74-.89)
- Classification errors differ for new (higher false positive rates) and established teams (higher false negative rates)
- Tentatively, best use for established teams with acceptable prior year fidelity scores
  - Screen positive \(\rightarrow\) Defer onsite for additional year
  - Screen negative \(\rightarrow\) Require onsite visit
Figure 1. A possible fidelity system

- New program? Yes -> Onsite interview
  No -> Score < 4.0?
  Yes -> Phone interview
  No -> Alarm event?
  Yes -> Phone interview
  No
Fig. 1 Conceptual model of implementation research

Fig. 1 Relationship between the EBSIS and the ISF. Solid lines indicate the original ISF (2008) figure and dashed lines indicate additions by our EBSIS approach. QA/QI are emphasized in two places: the provision of support to the Delivery System and the implementation of innovations (programs, policies, etc.)
Background—the good news: Explosion of interest in EBPs
The (potentially) bad news

- EBPs require fidelity monitoring to ensure accurate implementation
- The gold standard for fidelity monitoring is onsite which requires considerable assessment time for both assessor and agency
- The burden to the credentialing body, usually the state authority, increases exponentially with
  - The number of potential EBPs
  - The number of sites adopting each EBP
The problem may be worse than we think. Are there just 5 psychosocial EBPs?
Or, are there over 100?

<table>
<thead>
<tr>
<th>Date</th>
<th>Review source</th>
<th>Number of EBPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>Division 12 Taskforce</td>
<td>22 effective, 7 probable</td>
</tr>
<tr>
<td>1998</td>
<td>Treatments that Work</td>
<td>44 effective, 20 probable</td>
</tr>
<tr>
<td>2001</td>
<td>National EBP Project</td>
<td>6 effective</td>
</tr>
<tr>
<td>2001</td>
<td>Chambless, Annual Review of Psychology Article</td>
<td>108 effective or probable for adults; 37 for children</td>
</tr>
<tr>
<td>2005</td>
<td>What works for whom</td>
<td>31 effective, 28 probable</td>
</tr>
<tr>
<td>2007</td>
<td>Treatments that Work</td>
<td>69 effective, 73 probable</td>
</tr>
<tr>
<td>2008</td>
<td>SAMHSA Registry</td>
<td>38 w/ experimental support; 58 legacy programs</td>
</tr>
</tbody>
</table>
Alternative quality assurance mechanisms to alleviate the assessment burden*

- Use of shorter scales  (NOTE: both the newly revised DACTS and IPS scales are longer)
- Increase length of time between fidelity assessments
- Use of need-based vs. fixed interval schedules of assessment
- Use of alternative methods of assessment (e.g., self report, phone)

### Fidelity Assessment Variables

<table>
<thead>
<tr>
<th>Mode</th>
<th>Face-to-face, Phone, Self-report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection site</td>
<td>On-site</td>
</tr>
<tr>
<td></td>
<td>Off-site</td>
</tr>
<tr>
<td>Data collector</td>
<td>External—outside assessor</td>
</tr>
<tr>
<td></td>
<td>Agency affiliated—within agency, but outside the team</td>
</tr>
<tr>
<td></td>
<td>Internal—self assessment by team/program</td>
</tr>
<tr>
<td>Instrument</td>
<td>Full/partial/screen</td>
</tr>
<tr>
<td>Data source</td>
<td>EMR, chart review, self-report, observation</td>
</tr>
<tr>
<td>Informants</td>
<td>Team leader, full team, specific specialties (e.g., nurse), clients, significant others</td>
</tr>
<tr>
<td>Team variables</td>
<td>Size, location, years of operation, developmental status</td>
</tr>
</tbody>
</table>
Summary: Factors that may impact reliability and validity

- Phone interrater reliability
  - No apparent impact of rater
  - ICCs show small increase over time/with experience
- Validity—phone vs. onsite differences partly explicable by:
  - Level of item fidelity
  - Rater (ICCs, but not raw errors)
Future: Fidelity Outcome Training Model

1. Fidelity
2. Set goals
3. Training
4. Change in ACT behavior
5. Client Outcomes

The diagram illustrates a circular flow between Fidelity, Set goals, Training, Change in ACT behavior, and Client Outcomes, emphasizing the iterative process of setting and achieving outcomes.
The 17-year odyssey

- Priorities for research funding
- Peer review of grants
- Publication priorities and peer review
- Research synthesis
- Evidence-based medicine movement
- Guidelines for evidence-based practice
- Practice
  Funding: population needs, demands; local practice circumstances; professional discretion; credibility and fit of the evidence.

- Academic appointments, promotion, and tenure criteria
Classification

- How many categories – two groups, three groups?
- Which (sub)scales used to classify—total scale only?
- Cut scores? (4 assumed)
- Which error is more problematic (false positives, false negatives)?
  - Sensitivity, specificity, PPP, NPP?
- What is the criterion for validity of classification?
  - Onsite vs. clinical judgment?
  - Confusing operationalization of construct with construct (ACT=DACTS?)
Assessment – Continuous rating

- Are the (sub)scales interval?
  - Interval across all levels of scale (1 vs. 2 same as 4 vs. 5?)
- Sensitivity to change
- What subunits of scale are psychometrically sound/appropriate
  - Total scale vs. subscales
  - Individual items
Data Analysis: Comparing Methods

- Inter-rater reliability
  - Total and subscale scores for each rater
    - Intraclass Correlation Coefficient (ICC) between two raters of each fidelity method (consistency)
    - Mean and range of absolute value of differences between raters for each method (consensus)

- Validity
  - Total and subscale scores for each method
    - ICCs between methods (consistency)
    - Mean and range of absolute value of differences between methods (consensus)

- Sensitivity and specificity analysis
Self-Report Versus Phone Fidelity

- **Total DACTS**: Self-Report 4.12, Phone 4.21
- **Human Resources**: Self-Report 4.22, Phone 4.29
- **Organizational Boundaries**: Self-Report 4.53, Phone 4.56
- **Nature of Services**: Self-Report 3.72, Phone 3.87
Example: ACT dismantling studies

- Single case manager vs. Team approach
  - Team approach leads to more stable hospital reductions (Bond, Pensec et al., 1991)
- Low vs Hi Caseload ratios
  - Lower caseloads → better outcomes (Jerrell, 1999)
- Peer counselors vs. non-peer counselors
  - Mixed results

---

ACT: Will the real critical ingredients please stand up?

- Considerable overlap in ingredients identified using different methods
- Ingredients evolved over time (team size, composition, no discharge)
- Different perspectives/methods yield different ingredients (client vs expert)
- Different questions yield different ingredients (helpful/beneficial vs. critical)
Another concern: Feedback is not necessarily helpful

The good
- Fidelity reports can be powerful tools for guiding program improvements
- *Goal setting:* Giving focus to implementation efforts
- *Educational function:* Helping teams understand the practice
- *Political document:* Providing leadership with “cover” to make changes
- *Reinforcement:* Offering validation to teams achieving high fidelity

The problematic
- Leadership and teams do not always value reports (evaluation apprehension)
- Feedback must be provided in a timely fashion to be useful
- To be most useful, fidelity reports also must provide concrete action steps
Summary results: Phone Fidelity Assessment

- Acceptable interrater reliability
- Promising evidence of concurrent validity
  - Strong correlation with onsite (ICC)
  - Majority of programs classified within .10 scale points compared to onsite total DACTS
  - Raw error differences show little evidence of systematic bias (over- or under-estimates)

- Burden
  - Relatively high for site (however, lower than onsite and on par with good internal quality assurance process)
  - Relatively low for assessor
Limitations

- Quality of phone and self-report data may have been influenced by knowledge of subsequent onsite “audit”
- Predictive validity not assessed
- Small sample size
- Participant sites were volunteers (enthusiastic, conscientious)
- Limited to Indiana
- Limited to one EBP
Limitations

- Not all sites participated (16/24 of teams)
- Sites were previously certified ACT teams in one state
- Phone fidelity used as criterion fidelity measure