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

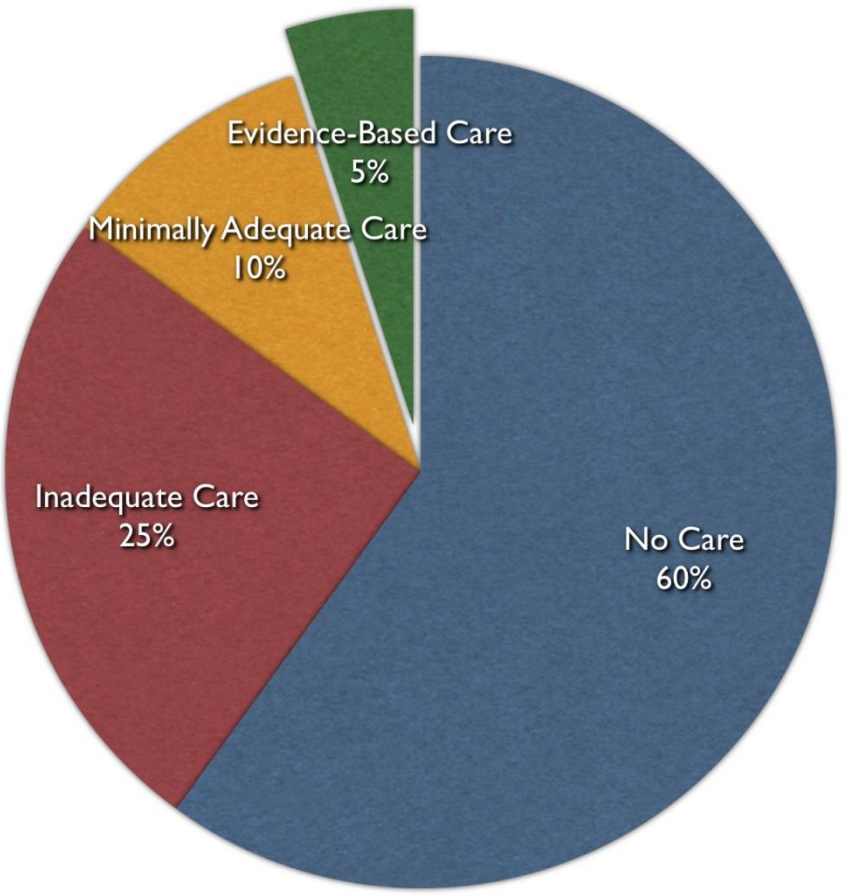
John McGrew, Ph.D.
Department of Psychology
Indiana University Purdue University Indianapolis

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)

1. President's New Freedom Commission on Mental Health , Achieving the Promise: Transforming Mental Health Care in America. Final Report. DHHS Pub. No. SMA-03-3832. Rockville, MD: 2003.
2. Institute of Medicine. “Improving the Quality of Health Care for Mental and Substance-Use Conditions: Quality Chasm 1 Series.” Washington: Institute of Medicine, November 2005.

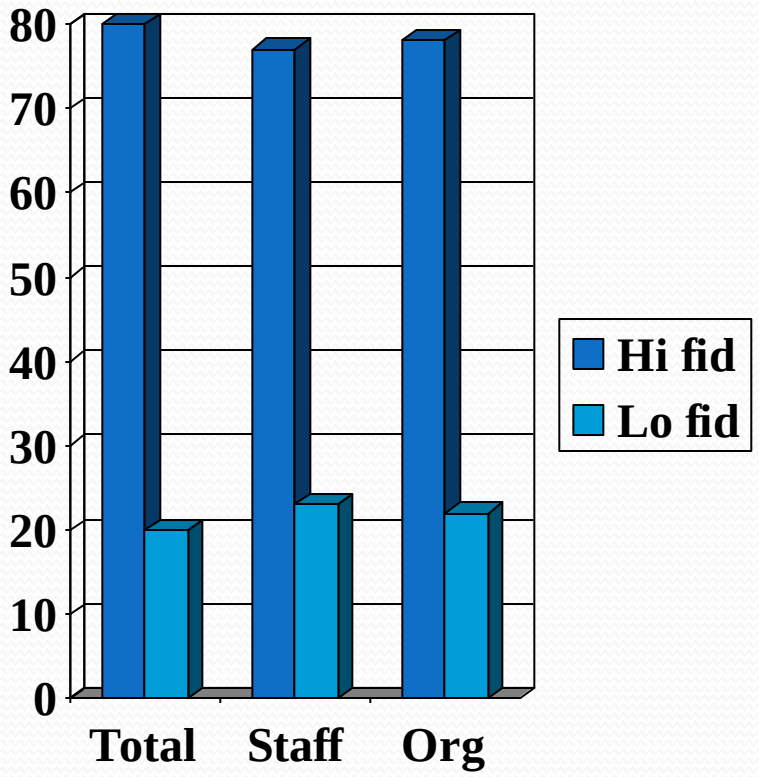
The implementation problem—It's probably Prozac

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

	ACT Compared to Controls		
	Better	No Diff.	Worse
Hospital use	17 (74%)	6 (26%)	0
Housing stability	8 (67%)	3 (25%)	1 (8%)
Symptoms	7 (44%)	9 (56%)	0
Quality of life	7 (58%)	5 (42%)	0

*__Source:__ Bond, GR, Drake, RE, Mueser, KT, & Latimer, E. (2001). Assertive Community Treatment for People with Severe Mental Illness. Dis Manage Health Outcomes, 9: 141-159.

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

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

Defining the model: Critical Organizational and Structural Ingredients

OK, we know our program works, but what exactly is working?

Critical ingredients: Some methodological issues

- Models elements usually defined BEFORE empirical testing→ 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

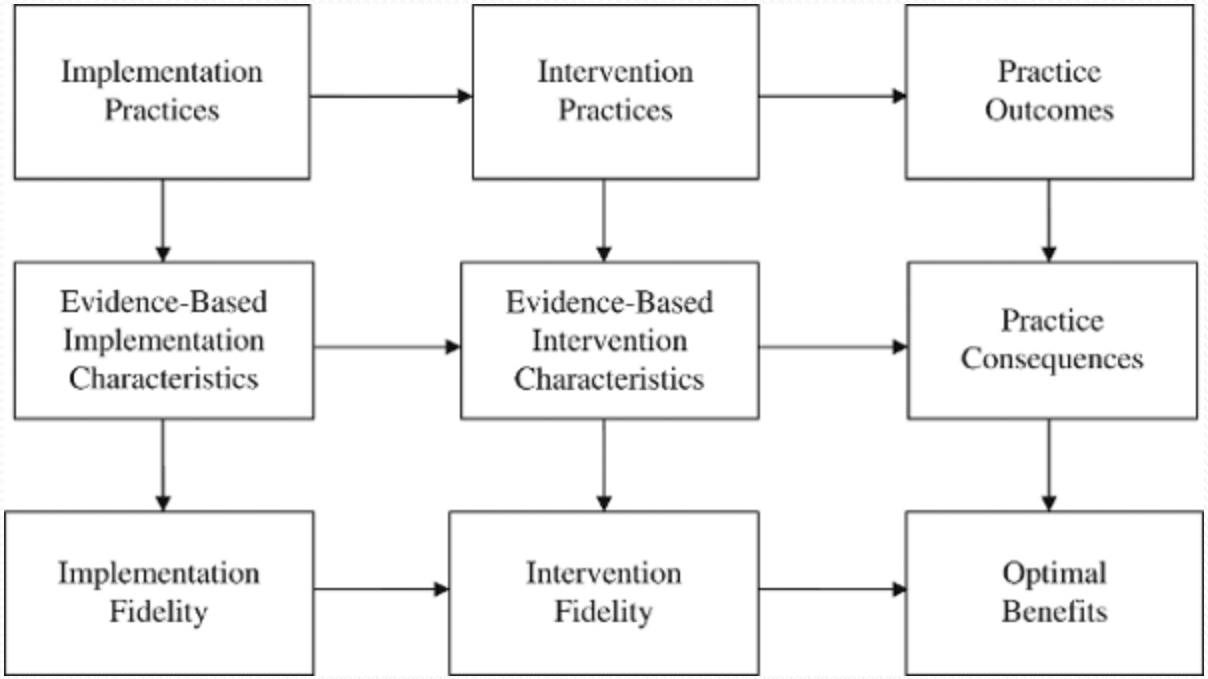
Shared caseloads	.65**
Number of contacts	.59**
24 hour availability	.55*
Daily team meeting	.49*
Nurse on team	.49*

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→ better outcomes (Jerrell, 1999)
- Peer counselors vs. non-peer counselors
  - Mixed results

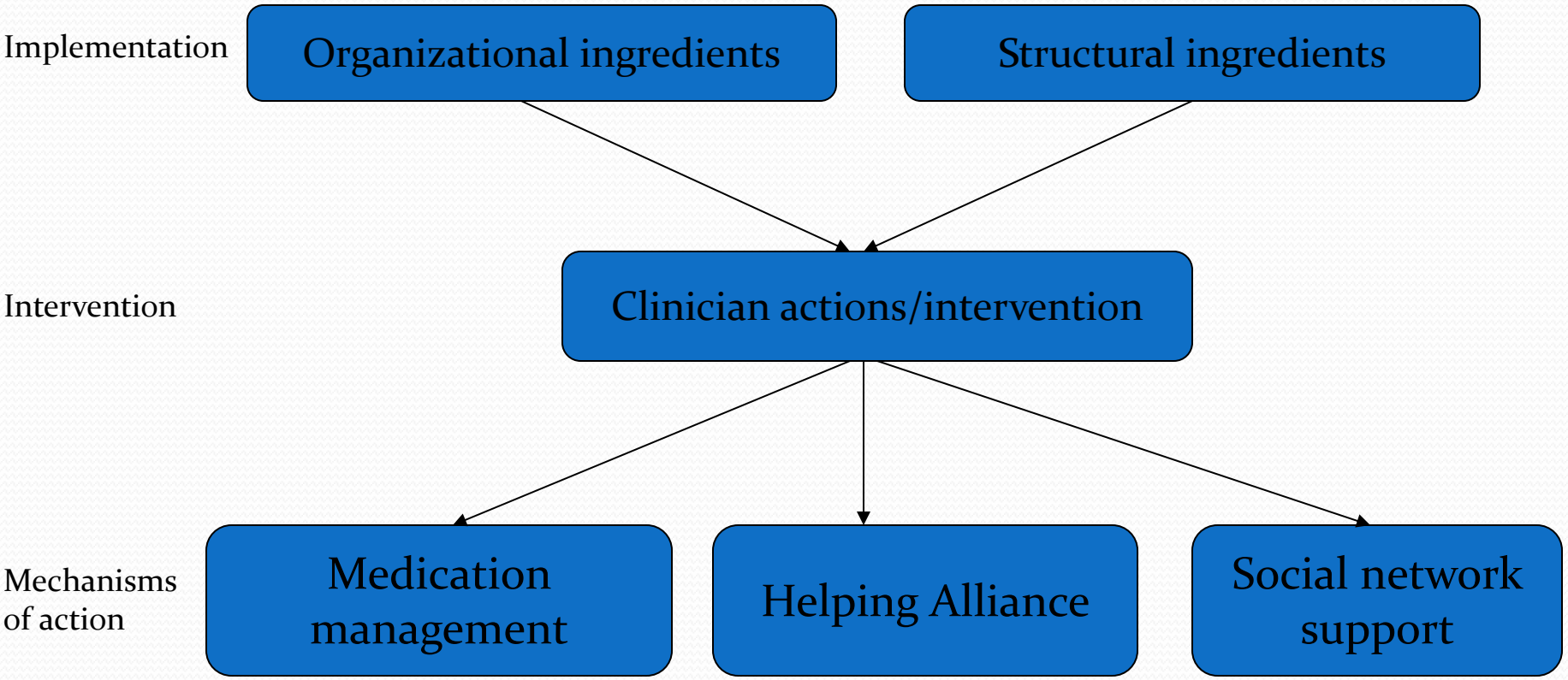
McGrew, J., Bond, G., Dietzen, L., & Salyers, M. (1994). Measuring the Fidelity of Implementation of a Mental Health Program Model. *Journal of Consulting and Clinical Psychology, 62*, 670-678.
McGrew, J. & Bond, G. (1997). The association between program characteristics and service delivery in Assertive Community Treatment. *Administration and Policy in Mental Health*, 25(2), 175-189.
Bond, G. R., Pensec, M., Dietzen, L., McCafferty, D., Giemza, R., & Sipple, H. W. (1991). Intensive case management for frequent users of psychiatric hospitals in a large city: A comparison of team and individual caseloads. Psychosocial Rehabilitation Journal, 15(1), 90-98.
Jerrell, J.M., & Ridgely, M.S. (1999). Impact of robustness of program implementation on outcomes of clients in dual diagnosis programs. *Psychiatric Services, 50*, 109–112.
Solomon, P., & Draine, J. (2001). The state of knowledge of the effectiveness of consumer provided services. Psychiatric Rehabilitation Journal, 25, 20-27.

Implementation vs. Intervention fidelity



Dunst, C.J. and C.M. Trivette, *Let's Be PALS: An Evidence-Based Approach to Professional Development.* Infants and Young Children, 2009. **22**(3): p. 164-176.

Inside the Black Box: a model of ACT helping



ACT workers' perspectives on clinical ingredients: Top ten ingredients

(N=73; McGrew et al., 2003)

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

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

	Successful (Fidelity >4)	Unsuccessful	Dropped Out
ACT	10 (77%)	3	
SE	8 (89%)	1	
IDDT	2 (15%)	9	2
IMR	6 (50%)	6	
FPE	3 (50%)	1	2
Total	29 (55%)	20	4

- EBPs differed in:
  - Clinical complexity
  - Practitioner familiarity
  - Compatibility with usual practice

Key difference: Type of fidelity items

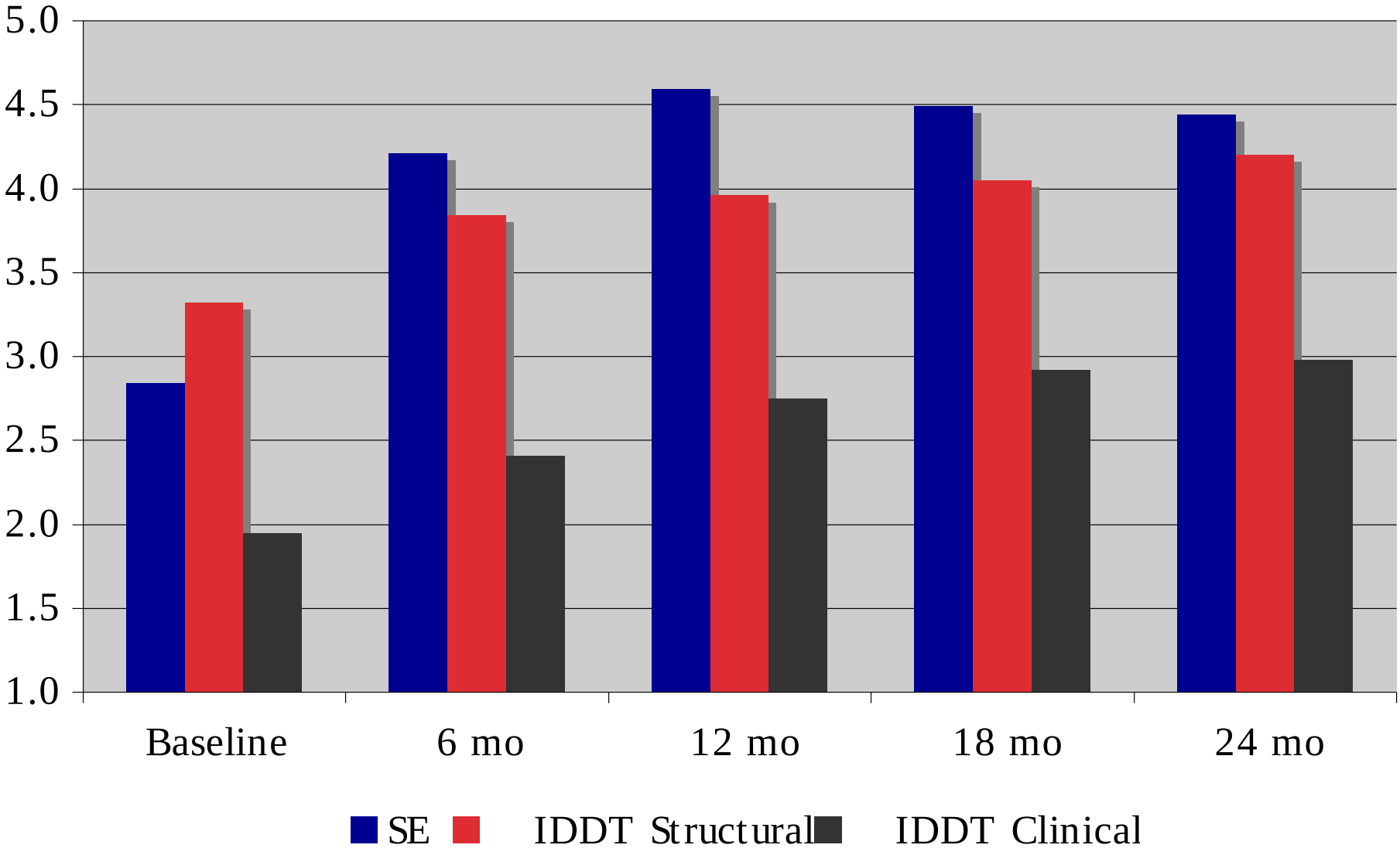
Structural Fidelity Items

- Things that can be done by administrative fiat, such as:
  - Daily team meetings
  - Multidisciplinary staffing
  - Low caseload ratio
  - Following a curriculum
  - Distributing educational handouts

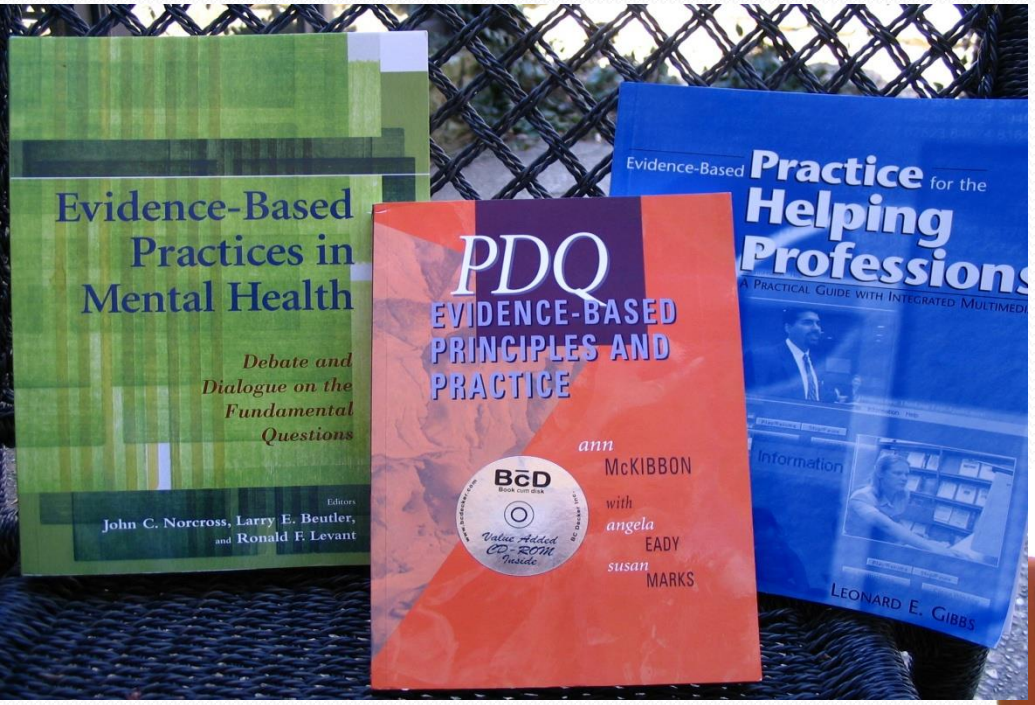
Assessing clinical interventions

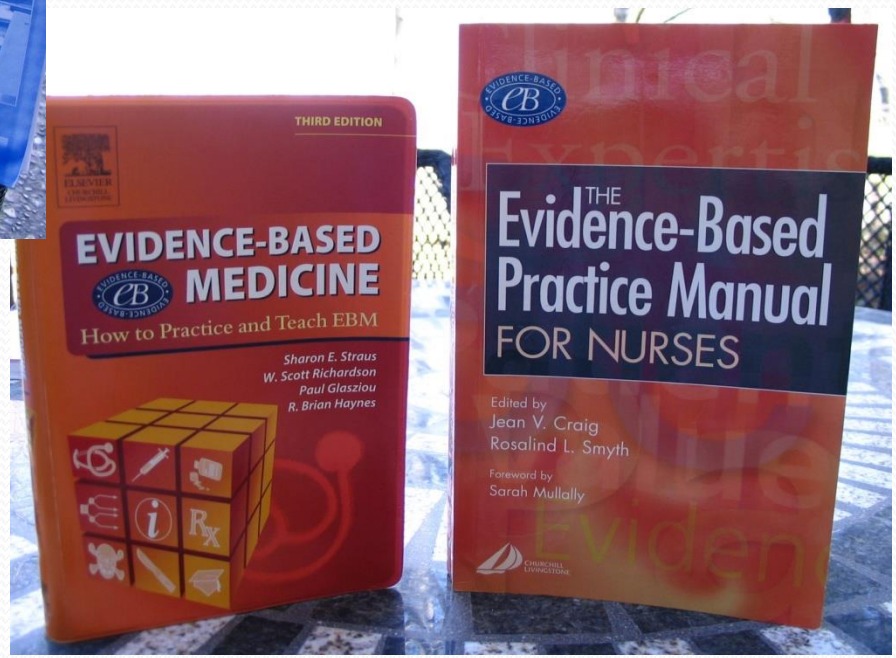
- Practitioner actions that follow prescribed techniques, such as:
  - Motivational interviewing
  - Behavioral tailoring
  - Providing stagewise interventions

Comparison of IDDT and SE Fidelity Over Time



Fidelity Burden—The elephant in the room: Explosion of interest in EBPs

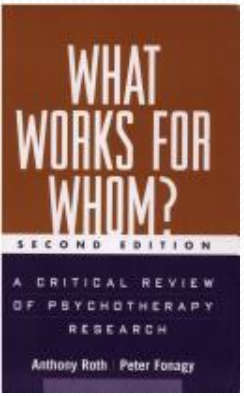




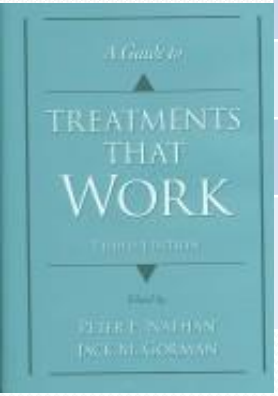
Current models for fidelity assessment are very time intensive

- 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



Date	Review source	Number of EBPs
1995	Division 12 Taskforce	22 effective, 7 probable
1998	Treatments that Work	44 effective, 20 probable
2001	National EBP Project	6 effective
2001	Chambless, Annual Review of Psychology Article	108 effective or probable for adults; 37 for children
2005	What works for whom	31 effective, 28 probable
2007	Treatments that Work	69 effective, 73 probable
2014	Division 12, APA	79 effective
2014	SAMHSA Registry	88 experimental, replicated programs



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)

*Evidence-based Practice Reporting for Uniform Reporting Service and National Outcome Measures Conference, Bethesda, Sept, 2007

Factors impacting fidelity assessment

Mode of collection	Face-to-face, Phone, Self-report
Designated rater	Independent rater, provider
Data collection site	On-site Off-site
Data collector	External—outside assessor Agency affiliated—within agency, but outside the team Internal—self assessment by team/program
Instrument	Full/ partial/ screen
Data source	EMR, chart review, self-report, observation
Informants	Team leader, full team, specific specialties (e.g., nurse), clients, significant others
Site variables potentially impacting	Size, location, years of operation, developmental status

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

Teague, G. B., Bond, G. R., & Drake .R.E. (1998). Program fidelity in assertive community treatment: development and use of a measure. *American Journal of Orthopsychiatry, 68*(2), 216-32.

DACTS Scoring

- Individual Items
  - Rating of ≤ 3 = Unacceptable implementation
  - Rating of 4 = Acceptable/good implementation
  - Rating of 5 = Excellent implementation
- Subscale scores and Total score
  - Mean of ≤4.0 = Below acceptable standards for adherence to model
  - Mean of 4.0-4.3 = Good adherence to model
  - Mean of ≥4.3 = Exemplary adherence to model

DACTS Items		Anchors				
Human Resources Items		1	2	3	4	5
H1	SMALL CASELOAD: client/provider ratio of 10:1.	50 clients/clinician or more.	35 - 49	21 - 34	11 - 20	10 clients/clinician or fewer
H2	TEAM APPROACH: Provider group functions as team rather than as individual practitioners; clinicians know and work with all clients.	Fewer than 10% clients with multiple staff face-to-face contacts in 2-weeks	10 - 36%.	37 - 63%.	64 - 89%.	90% or more clients have face-to-face contact with > 1 staff member in 2 weeks.
H3	PROGRAM MEETING: Program meets frequently to plan and review services for each client.	Program service-planning for each client usually occurs once/month or less frequently.	At least twice/month but less often than once/week.	At least once/week but less often than twice/week.	At least twice/week but less often than 4 times/week.	Program meets at least 4 days/week and reviews each client each time, even if only briefly.
H4	PRACTICING TEAM LEADER: Supervisor of front line clinicians provides direct services.	Supervisor provides no services.	Supervisor provides services on rare occasions as backup.	Supervisor provides services routinely as backup, or less than 25% of the time.	Supervisor normally provides services between 25% and 50% time.	Supervisor provides services at least 50% time.
H5	CONTINUITY OF STAFFING: program maintains same staffing over time.	Greater than 80% turnover in 2 years.	60-80% turnover in 2 years.	40-59% turnover in 2 years.	20-39% turnover in 2 years.	Less than 20% turnover in 2 years.
H6	STAFF CAPACITY: Program operates at full staffing.	Program has operated at less than 50% of staffing in past 12 months.	50-64%	65-79%	80-94%	Program has operated at 95% or more of full staffing in past 12 months.
H7	PSYCHIATRIST ON STAFF: there is at least one full-time psychiatrist per 100 clients assigned to work with the program.	Program for 100 clients has less than .10 FTE regular psychiatrist.	.10-.39 FTE per 100 clients.	.40-.69 FTE per 100 clients.	.70-.99 FTE per 100 clients.	At least one full-time psychiatrist is assigned directly to a 100-client program.

Study 1: Phone Based Assessment

Why phone based?
Preliminary studies demonstrating predictive validity

	Correlations between closure rates and total fidelity scores in Supported Employment	
	QSEIS and VR closure rates	IPS and VR closure rates
McGrew & Griss, 2005, n=23	.42*	-.07
McGrew, 2007, n=17	n/a	.37t
McGrew, 2008, n=23	n/a	.39*

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?

McGrew, J., Stull, L., Rollins, A., Salyers, M., & Hicks, L. (2011). A comparison of phone-based and onsite-based fidelity for Assertive Community Treatment (ACT): A pilot study in Indiana. *Psychiatric Services, 62,* 670-674

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?

Format used for phone protocol

Client	**Admission – team involved?**	**Discharge – team involved?**
Example	*Team brought client into ER and helped with inpatient admission documentation*	*Team participated in discharge planning prior to release, transported him home upon release*
Client 1		
Client 2		
Client 3		
Client 4		
Client 5		
Client 6		
Client 7		
Client 8		
Client 9		
Client 10		

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.***

	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)
Living in supervised living situation	
Other housing support outside the ACT team	
Psychiatric services	
Case management	
Counseling/ individual supportive therapy	
Substance abuse treatment	
Employment services	
Other rehabilitative services	

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

Comparison – total DACTS scores	Single Measure ICC	Average Measure ICC
Experienced rater vs. second rater	0.91	0.93
ONSITE published estimate* Comparing consultant, trainer and implementation monitor	0.99[1]	

*McHugo, G.J., Drake, R.E., Whitley, R., Bond, G.R., et al. (2007). Fidelity outcomes in the national implementing evidence-based practices project. *Psychiatric Services, 58*(10), 1279-1284.

Note 1. Type of ICC not specified

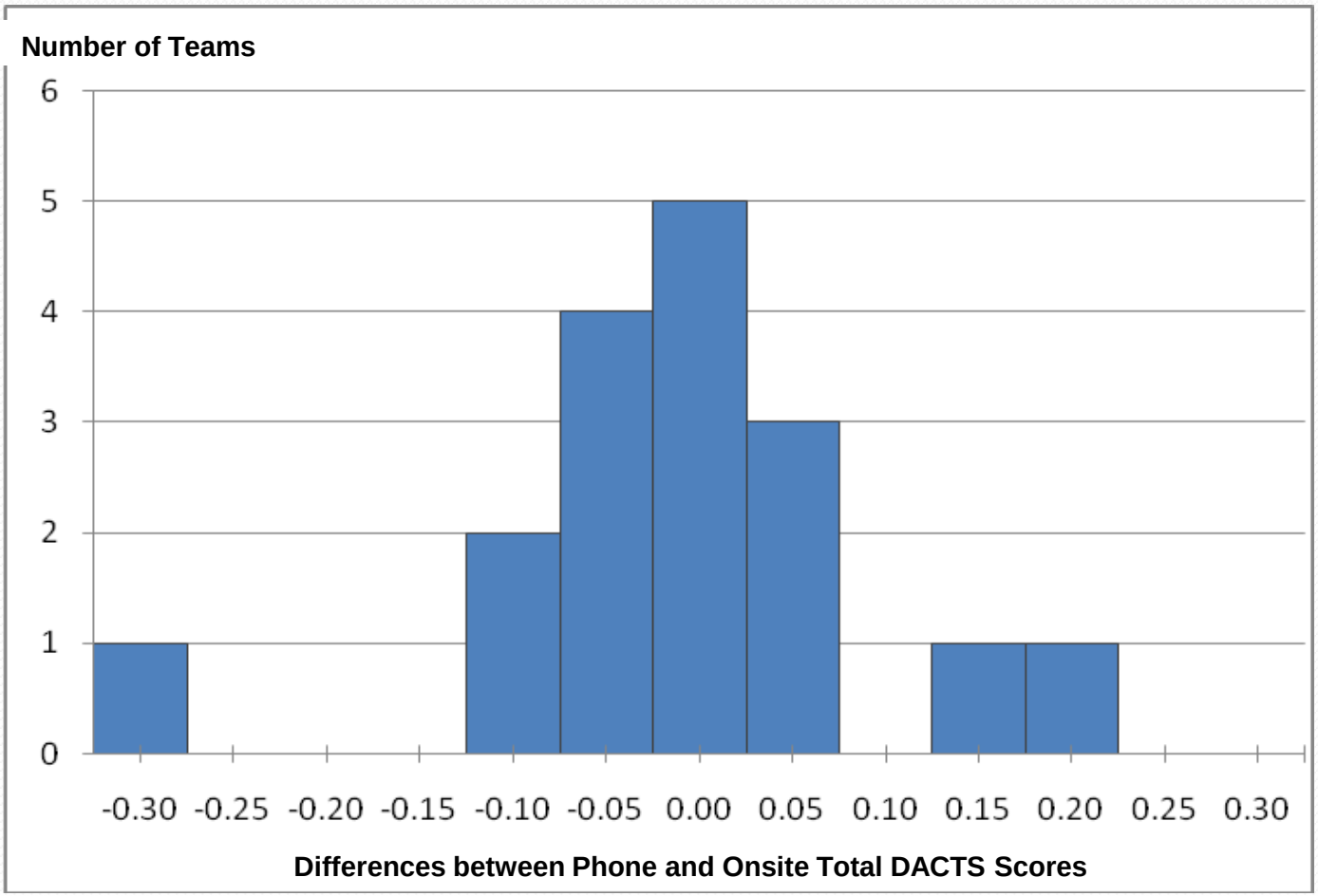
Results: Phone based assessment is valid compared to onsite (consistency)

Comparisons using DACTS Total Score	Single Measures ICC	Average Measures ICC
Onsite vs. Phone Consensus	0.87	0.93

Phone based had adequate validity compared to onsite for total and subscale scores (consensus)

Item/Subscale	Phone Consensus Mean/SD (n = 17)	Onsite Mean/SD (n = 17)	Mean Absolute Difference (n = 17)	Range of Absolute Differences	Intraclass Correlation Coefficients
Total DACTS	4.29 (0.19)	4.30 (0.13)	0.07	0.00 – 0.32	0.87
Organizational Boundaries	4.72 (0.19)	4.74 (0.18)	0.08	0.00 – 0.29	0.73
Human Resources	4.35 (0.22)	4.34 (0.28)	0.12	0.00 – 0.27	0.87
Services	3.91 (0.31)	3.95 (0.23)	0.14	0.00 – 0.50	0.86

Frequency distribution of differences between onsite and phone total DACTS scores



DACTS Phone Assessment Burden

Task	Time (Mean/SD)	Time Range
Site Preparation for call	7.5 hours (6.2)	1.75 to 25
Phone call	72.8 minutes (18.5)	40 to 111

Explaining the results: Reliability tends to improve over time

Comparisons using DACTS Total Score	Single Measures ICC
Experienced vs. Second rater (1st 8 sites)	0.88
Experienced vs. Second rater (Last 9 sites)	0.95

Explaining the differences:
Rater expertness or prior experience with the site does not influence interrater *reliability*

Comparison	Experienced Phone M/SD	Comparison Rater Phone M/SD	Mean Absolute Difference	Range of Absolute Differences	ICC
Experienced vs. Rater 2	4.29 (0.18)	4.31 (0.19)	0.06	0.00 – 0.25	0.91
Experienced vs. Trainer	4.38 (0.14)	4.44 (0.14)	0.08	0.00 – 0.25	0.92
Experienced vs. Naïve	4.21 (0.19)	4.19 (0.16)	0.05	0.00 – 0.14	0.91

Explaining the differences:
Rater prior experience/expertness may influence concurrent *validity* (consistency, but not consensus)

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

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

McGrew, J., White, L., Stull, L., & Wright-Berryman, J. (2013). A comparison of self-reported and phone-based fidelity for Assertive Community Treatment (ACT): A pilot study in Indiana. *Psychiatric Services*. Published online January 3, 2013.

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)

Reliability comparisons (n=16)	Experienced Rater		Naïve Rater		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
Total DACTS (Experienced vs. Second Rater)	**4.22**	.25	**4.20**	.28	**.04**	.00 – 0.11	.98
Organizational Bound. Subscale	**4.58**	.14	**4.57**	.14	**.06**	.00 – 0.14	.77
Human Resources Subscale	**4.27**	.35	**4.30**	.36	**.05**	.00 – 0.27	.97
Nature of Services Subscale	**3.91**	.41	**3.84**	.46	**.07**	.00 – 0.40	.97

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

Reliability comparisons (n=16)	Consultant Rater		Experienced Rater		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
Total DACTS	**4.16**	.27	**4.11**	.26	**.14**	.00 – 0.41	**.77**
Organizational Bound. Subscale	**4.49**	.20	**4.53**	.21	**.13**	.00 – 0.42	**.61**
Human Resources Subscale	**4.27**	.39	**4.21**	.28	**.25**	.00 – 0.91	**.47**
Nature of Services Subscale	**3.72**	.50	**3.76**	.48	**.20**	.00 – 0.60	**.86**

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)

Validity comparisons (n=16)	Self-Report		Phone		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
Total DACTS	**4.12**	.27	**4.21**	.27	**.13**	.00 - .43	**.86**
Organizational Bound. Subscale	**4.53**	.15	**4.56**	.12	**.08**	.00 - .29	**.71**
Human Resources Subscale	**4.22**	.31	**4.29**	.34	**.15**	.00 – 64	**.74**
Nature of Services Subscale	**3.72**	.49	**3.87**	.47	**.20**	.07 - .50	**.92**

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)

Items	Subscale	Self-Report	Phone	Difference	Significance
Dual Diagnosis Model	Nature of Services	3.80	4.56	**.76**	t = 4.58 **p < .001**
Vocational Specialist	Human Resources	3.25	3.88	**.63**	t = 1.67 p = .116
Informal Support System	Nature of Services	3.00	3.44	**.44**	t = 1.60 p = .130
Responsibility for Crisis Services	Organizational Boundaries	4.31	4.69	**.38**	t = 3.00 **p = .009**
Consumer on Team	Nature of Services	1.75	1.38	**.37**	t = -1.38 p = .189
Responsibility for Tx Services	Organizational Boundaries	4.44	4.69	**.25**	t = 2.23 p = .041
Continuity of Staff	Human Resources	3.31	3.06	**.25**	t = 1.379 p = .188

Classification: Sensitivity and Specificity

ACT Team = Fidelity Score ≥ 4.0, Phone=criterion

		Phone		
		ACT Team	Not ACT Team	Total
Self-Report	ACT Team	**10**	0	10
	Not ACT Team	3	**3**	6
	Total	13	3	**16**

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 (≤ .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

Validity comparisons (n=32)	Onsite		Phone		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
Total DACTS	**3.22**	.28	**3.15**	.28	**.13**	.00 – 0.50	**.88**
Organizational Bound. Subscale	**3.76**	.38	**3.64**	.35	**.18**	.00 – 0.80	**.85**
Human Resources Subscale	**3.38**	.41	**3.35**	.43	**.16**	.00 – 0.70	**.94**
Nature of Services Subscale	**2.66**	.33	**2.60**	.31	**.18**	.00 – 0.70	**.84**

Validity of self-report vs. onsite is good to acceptable

Validity comparisons (n=32)	Onsite		Self-report		Mean Absolute Difference	Range of Absolute Differences	Intraclass Correlation Coefficient
	Mean	SD	Mean	SD			
Total DACTS	**3.22**	.28	**3.17**	.31	**.17**	.00 – 0.60	**.84**
Organizational Bound. Subscale	**3.76**	.38	**3.62**	.40	**.26**	.00 – 1.3	**.66**
Human Resources Subscale	**3.38**	.41	**3.35**	.48	**.19**	.00 – .50	**.92**
Nature of Services Subscale	**2.66**	.33	**2.66**	.40	**.25**	.00 – 0.70	**.79**

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

1. Bond, G. (2013) Self-assessed fidelity: Proceed with caution. *Psychiatric Services, 64*(4), 393-4.
2. McGrew, J.H., White, L.M., & Stull, L. G. (2013). Self-assessed fidelity: Proceed with caution: In reply. *Psychiatric Services, 64*(4), 394

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

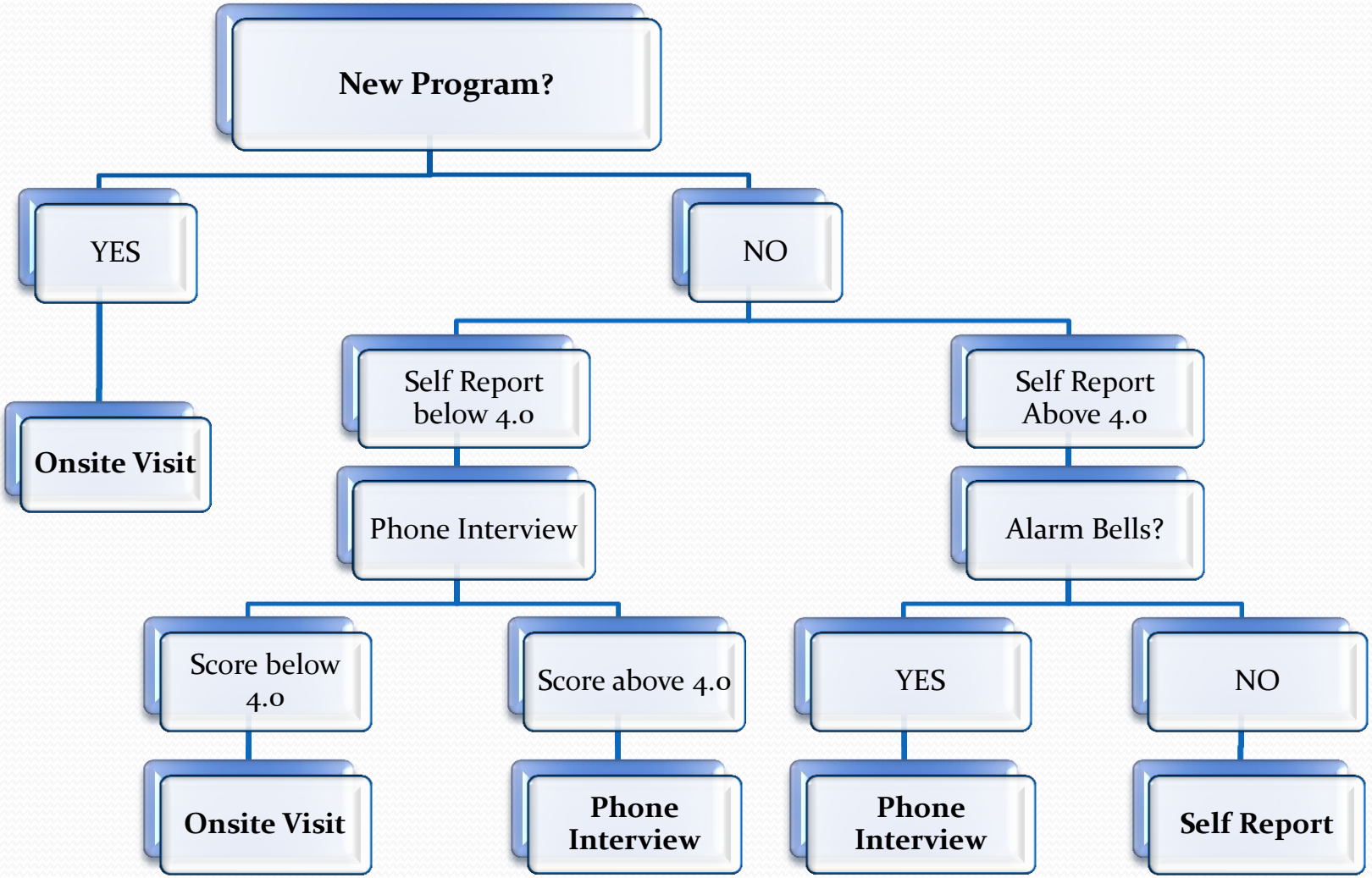
McGrew, J. (2011). The TMACT: Evidence based or eminence based? *Journal of the American Psychiatric Nursing Association, 17*, 32-33. (letter to the editor)

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

1. McGrew, J., Stull, L., Rollins, A., Salyers, M., & Hicks, L. (2011). A comparison of phone-based and onsite-based fidelity for Assertive Community Treatment (ACT): A pilot study in Indiana. *Psychiatric Services, 62,* 670-674
2. McGrew, J. H., & Stull, L. (September 23, 2009). Alternate methods for fidelity assessment. Gary Bond Festschrift Conference, Indianapolis, IN

Fidelity Assessment System



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

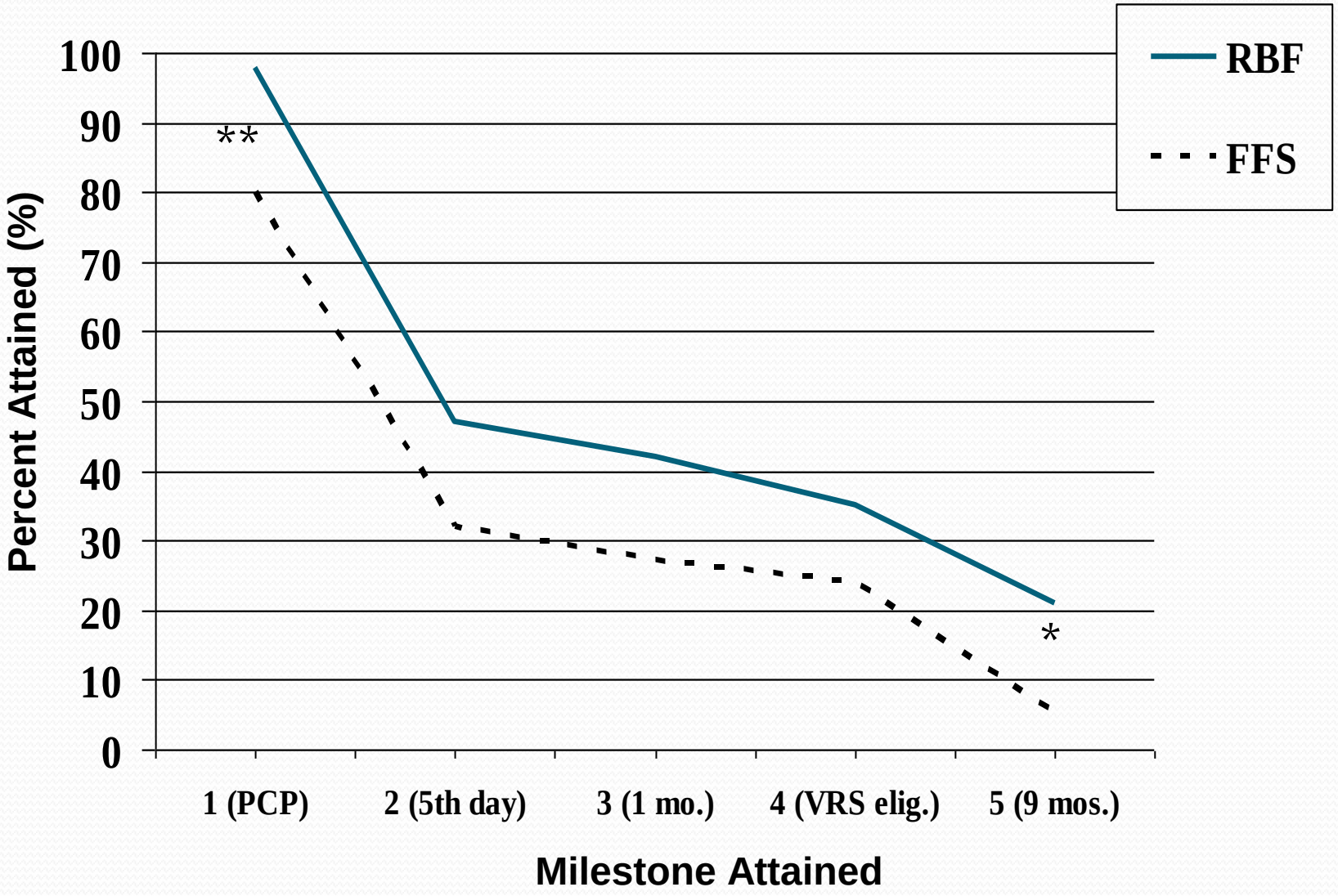
Monroe-Devita et al. (2012). Program fidelity and beyond: Multiple strategies and criteria for ensuring quality of Assertive Community Treatment. Psychiatric Services, 63, 743-750.

An alternate to fidelity

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

McGrew, J.H, Johannesen, J.K., Griss, M.E., Born, D., & Hart Katuin, C. (2005). Performance-based funding of supported-employment: A multi-site controlled trial. *Journal of Vocational Rehabilitation, 23*, 81-99.
McGrew, J.H, Johannesen, J.K., Griss, M.E., Born, D., & Hart Katuin, C. (2007) Performance-based funding of supported employment: Vocational Rehabilitation and Employment staff perspectives. *Journal of Behavioral Health Services Research, 34,* 1-16.
McGrew, J., Newman, F., & DeLiberty, R. (2007). The HAPI-Adult: The Psychometric Properties of an Assessment Instrument Used to Support Service Eligibility and Level of Risk-Adjusted Reimbursement Decisions in a State Managed Care Mental Health Program. *Community Mental Health Journal,43,*481-515.

Results Based Funding: Milestone Attainment Across Sites



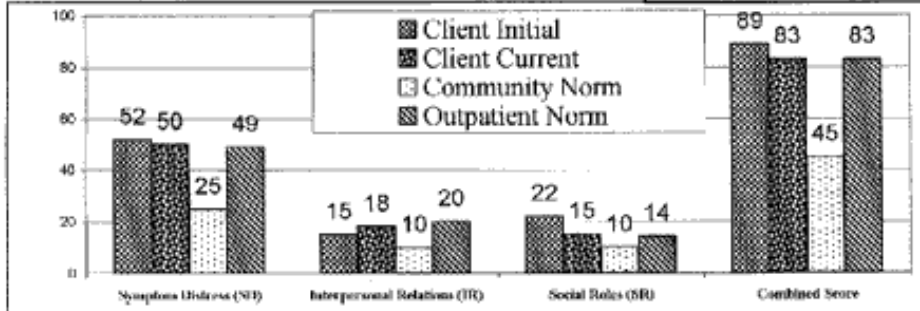
*p < .05, **p < .01

Performance tracking

Dr. John Smith
Session #13

Client:	12345678
Provider:	0
Date:	8/18/99
Time:	9:14 PM
Duration (mins):	10.74
Avg call (mins):	6.2

OQ-45.2



Client answered 45 of 45 questions.

Progress: GREEN: The rate of change the patient is making is in the adequate range. No change in treatment plan is recommended on the results.

Symptom Distress This scale measures subjective discomfort related to symptoms of anxiety and depression.

Interpersonal Relations This scale measures friction, conflict, inadequacy and withdrawal in friendships, family and marriage.

Social Role This scale measures dissatisfaction, conflict, distress and inadequacy in performance of tasks related to employment, school, family roles and leisure life.

OQ-45.2 Critical Questions	Answer
8. I have thoughts of ending my life.	Rarely
23. I feel hopeless about the future.	Frequently
26. I feel annoyed by people who criticize my drinking/drug use.	Never
32. I have trouble at work or school because of drinking or drug use.	Never
44. I feel angry enough at work or school to do something I might regret.	Rarely

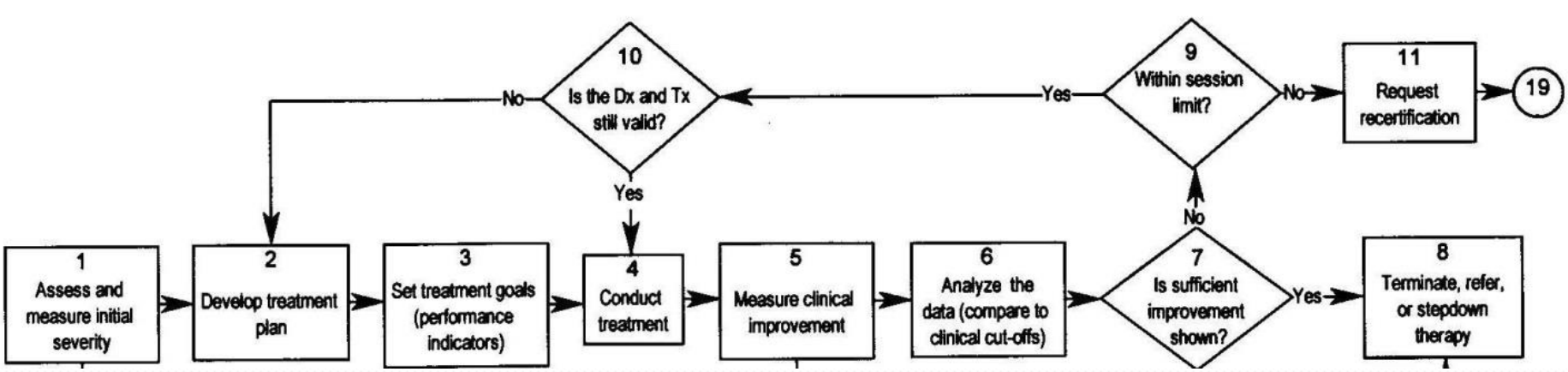
Client Extreme Responses	Answer
15. I feel worthless.	Almost Always
18. I feel lonely.	Almost Always
22. I have difficulty concentrating.	Almost Always

OQ™ Systems, Inc.
34 Wall Street, Suite 6
Norwalk, CT 06850
800/ 357 1200
www.oqsystems.com

TELESAGE™
The Wise Choice in Survey Systems

TeleSage
4558 Fourth Avenue NE
Seattle, WA 98105-4813
800-636-8524
www.telesage.com

Alternate to fidelity: Outcome management



Lambert, M. et al. (2000). Quality improvement: Current research in outcome management. In G. Stricker, W. Troy, & S. Shueman (eds). Handbook of Quality Management in Behavioral Health (pp. 95-110). Kluwer Academic/Plenum Publishes, New York

Thanks to the following collaborators!

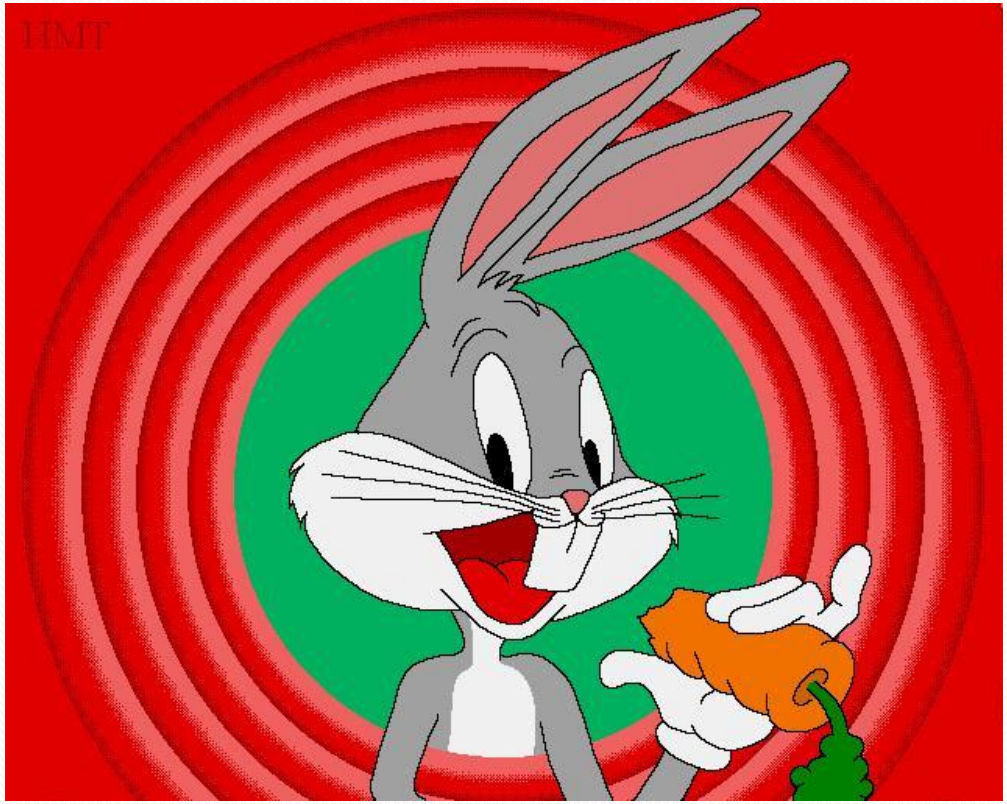
- Angie Rollins
- Michelle Salyers
- Alan McGuire
- Lia Hicks
- Hea-Won Kim
- David McClow
- Jennifer Wright-Berryman
- Laura Stull
- Laura White

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?

	Pearson Correlation
Human Resources Subscale	-0.83**
Organizational Boundaries Subscale	-0.67**
Services Subscale	-0.58* (0.27)[1]
Total DACTS	-0.74** (-0.34)[1]

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

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

Note 1: includes S10–peer specialist

Phone Fidelity

Strengths

- Strong Reliability
- Strong validity with onsite visit[16]
- Less burdensome than onsite visit
- Gathers more detailed information than self-report
- Identifies missing data
- Personal communication with TL (and other members of team)
- Opportunity to discuss issues, problems, feedback, etc.

Weaknesses

- Time intensive
- Scheduling issues
- Less comprehensive than onsite fidelity visit
- May be redundant with self-report fidelity

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

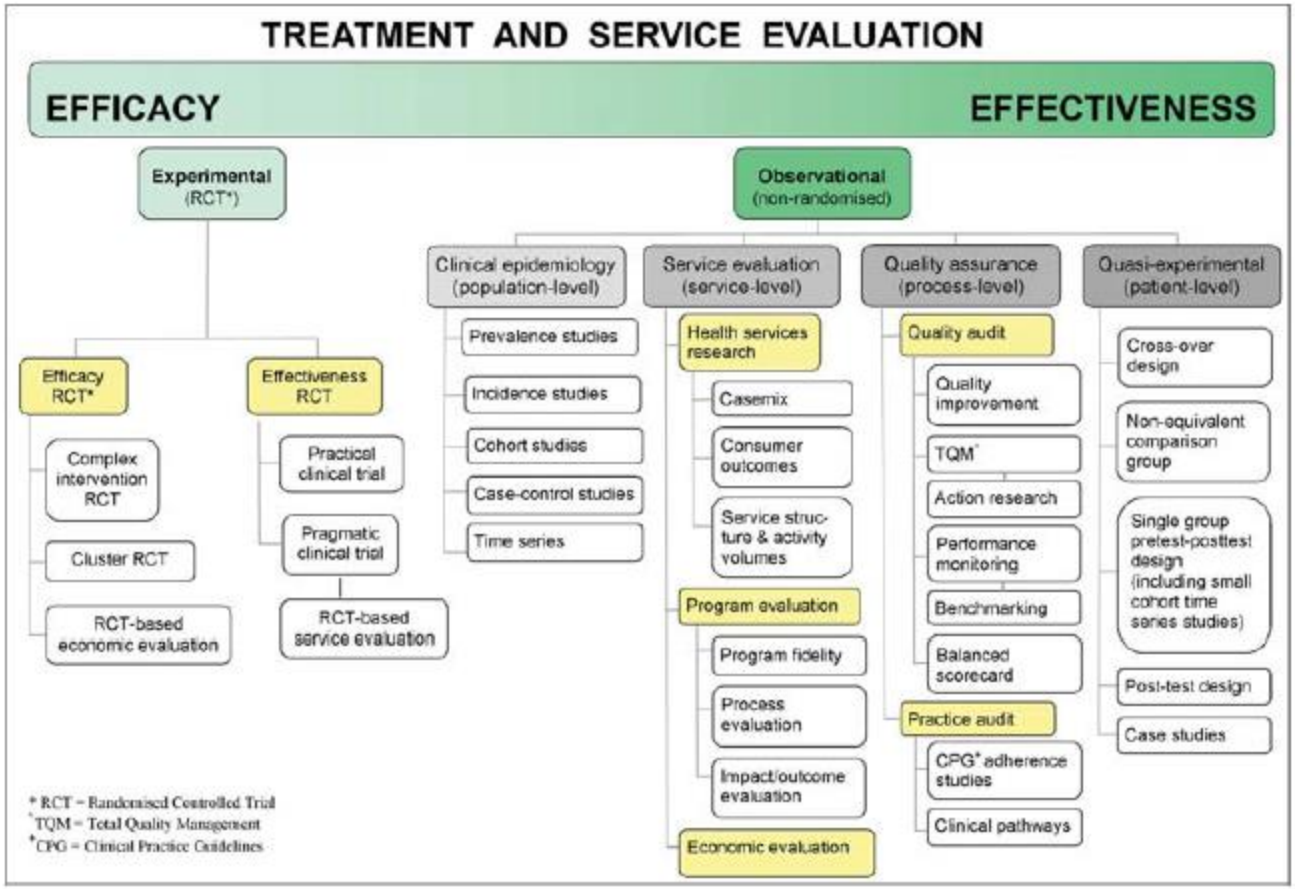


Figure 1. Treatment and service evaluation. CPG, clinical practice guidelines; RCT, randomized controlled trial; TQM, total quality management.

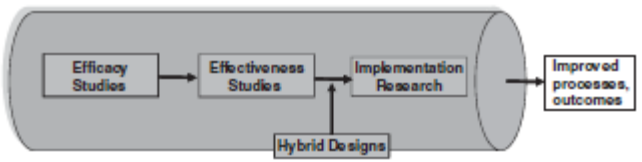
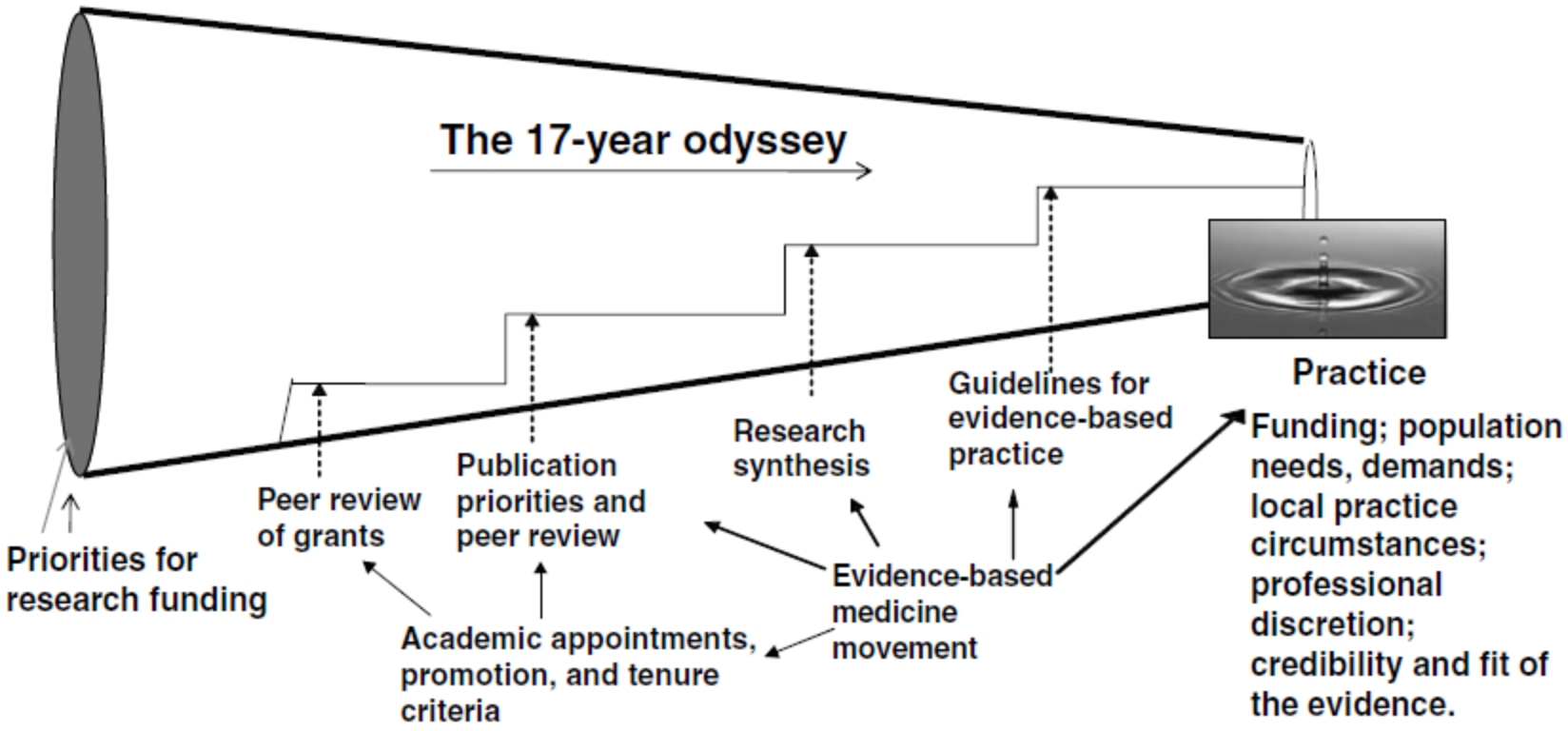
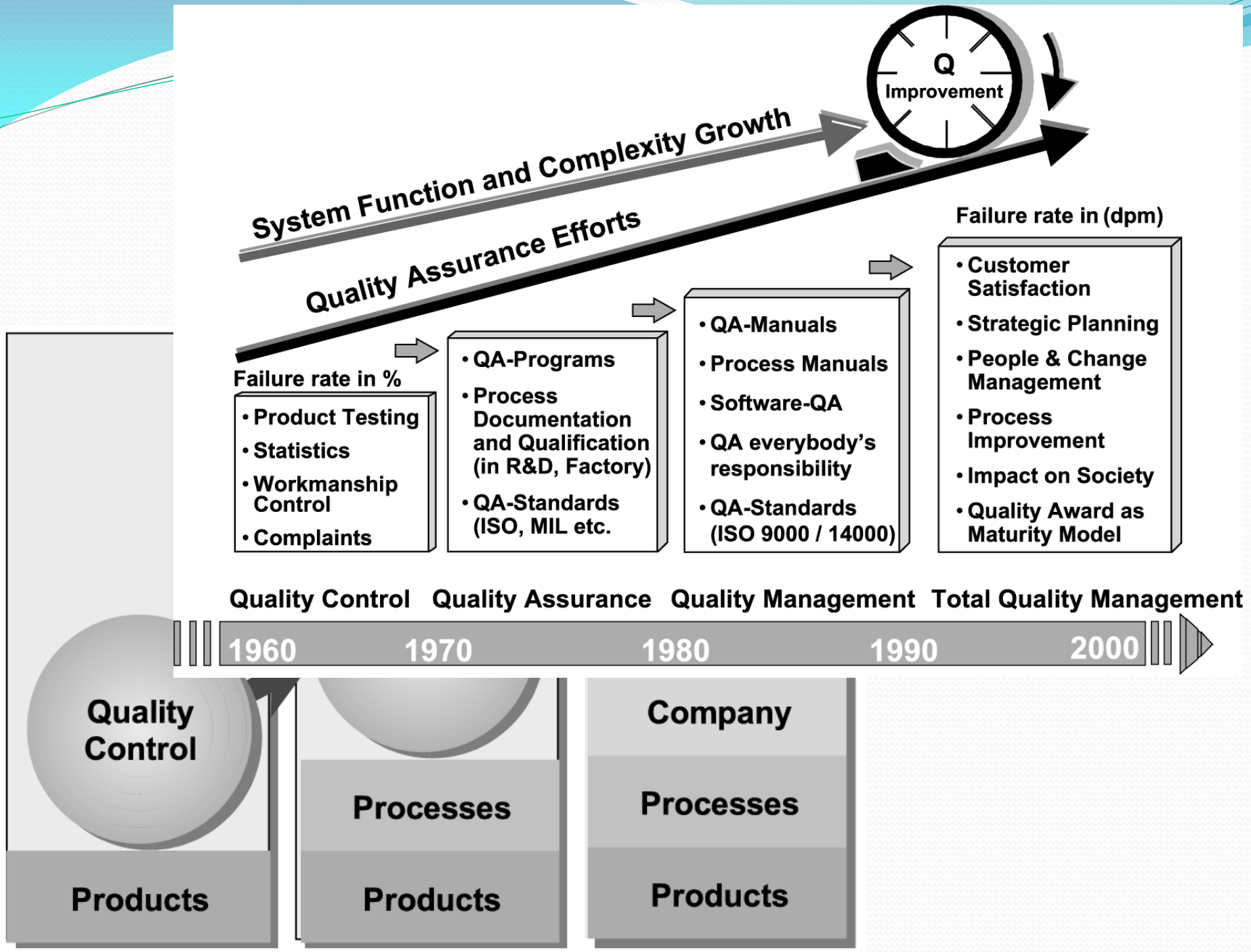
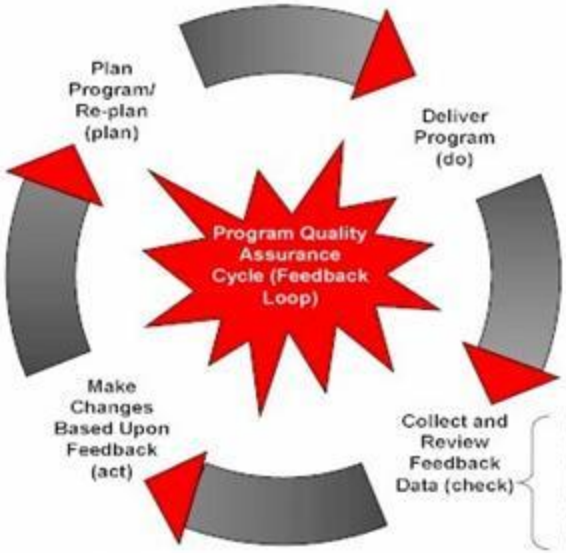


FIGURE 1. Research pipeline.




System Function and Complexity Growth
Quality Assurance Efforts
Q
Improvement
Failure rate in %
• Product Testing
• Statistics
• Workmanship Control
• Complaints
• QA-Programs
• Process Documentation and Qualification (in R&D, Factory)
• QA-Standards (ISO, MIL etc.
• QA-Manuals
• Process Manuals
• Software-QA
• QA everybody's responsibility
• QA-Standards (ISO 9000 / 14000)
Failure rate in (dpm)
• Customer Satisfaction
• Strategic Planning
• People & Change Management
• Process Improvement
• Impact on Society
• Quality Award as Maturity Model
Quality Control
Quality Assurance
Quality Management
Total Quality Management
1960
1970
1980
1990
2000
Quality Control
Products
Processes
Products
Company
Processes
Products

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

		Discrimination between ACT, ICM and BRK (F-test) n=87	Item total (mean r across 3 years) ACCESS sites n=18	Item total (16-state) n=35	Item total (ACT Center baseline) n=31	Times in top-10
H1	Small caseload	29.6		0.62	0.46	3
H2	team approach	14.9		0.55		2
H3	Program meeting					0
H4	Practicing Leader		0.43		0.32	2
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			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	0.64	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	0.49	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			

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)

		DACTS Total Score					
		16 State		ACT Center Baseline		ACT Center Follow-up	
		ACT	Non-ACT	ACT	Non-ACT	ACT	Non-ACT
DACTS screen	ACT	7	3	9	8	81	7
	Non-ACT	1	24	0	14	8	17
Correlation with DACTS		.86		.86		.83	
Sensitivity		.88		1.0		.91	
Specificity		.89		.64		.71	
PPP		.70		.53		.92	
NPP		.96		1.0		.68	
Overall PP		.89		.74		.87	

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 → Defer onsite for additional year
  - Screen negative → Require onsite visit

Figure 1. A possible fidelity system

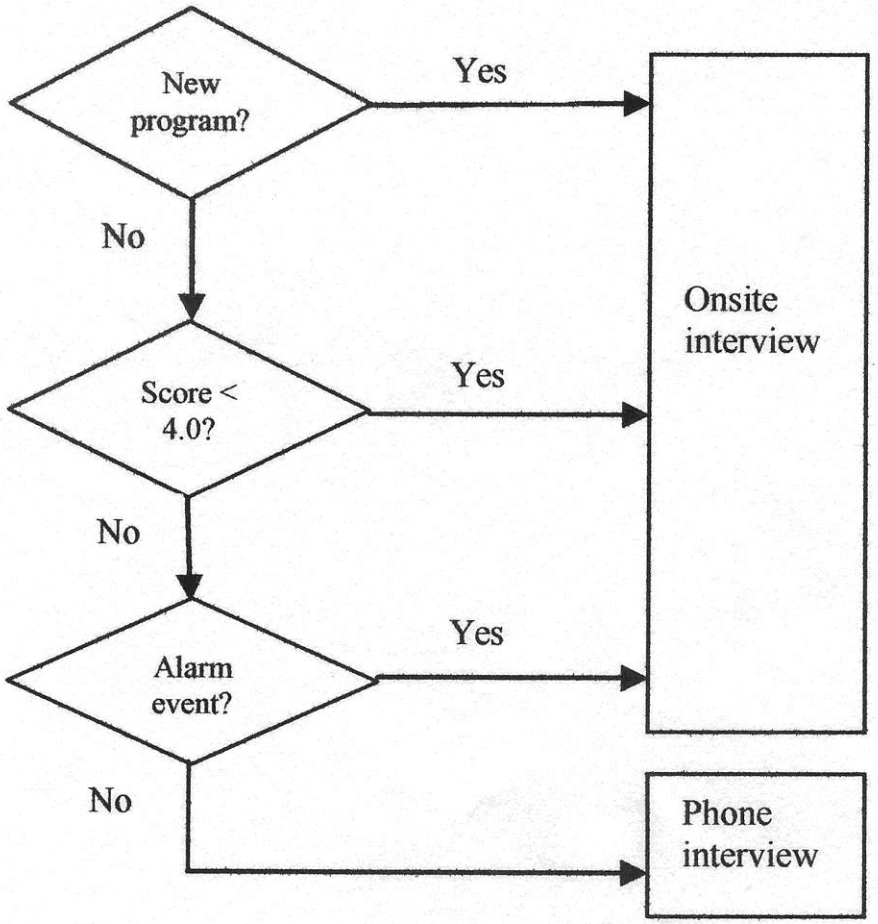
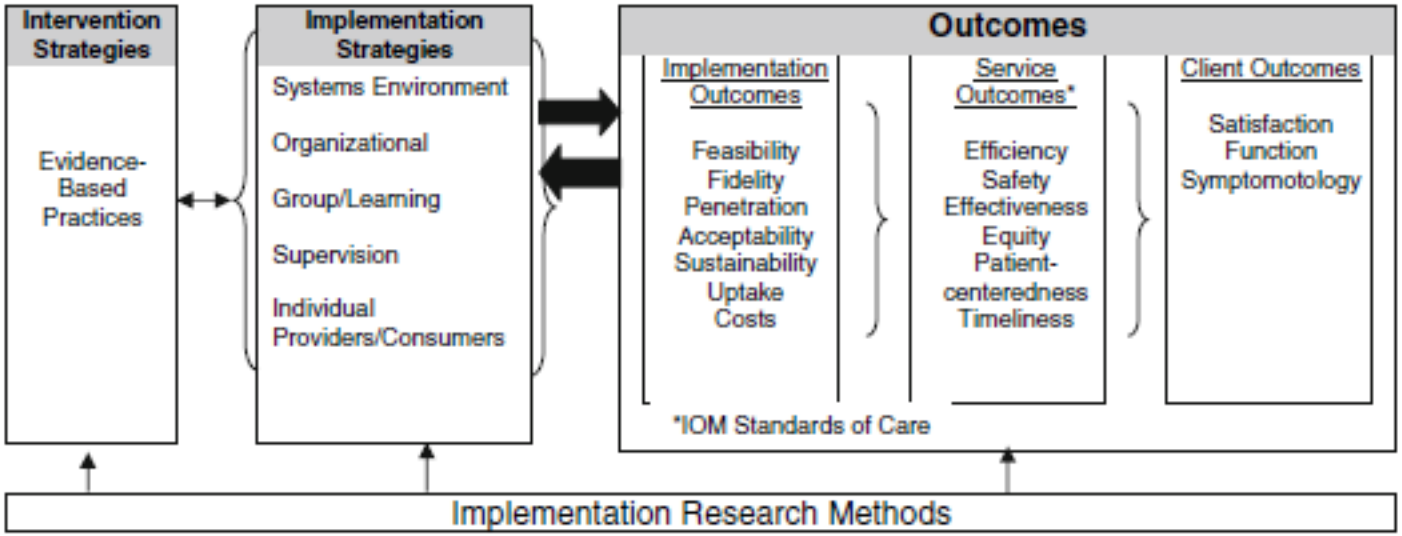


Fig. 1 Conceptual model of implementation research



Proctor, et al. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological and training challenges. Administration and Policy in Mental Health, 36, 24-34.

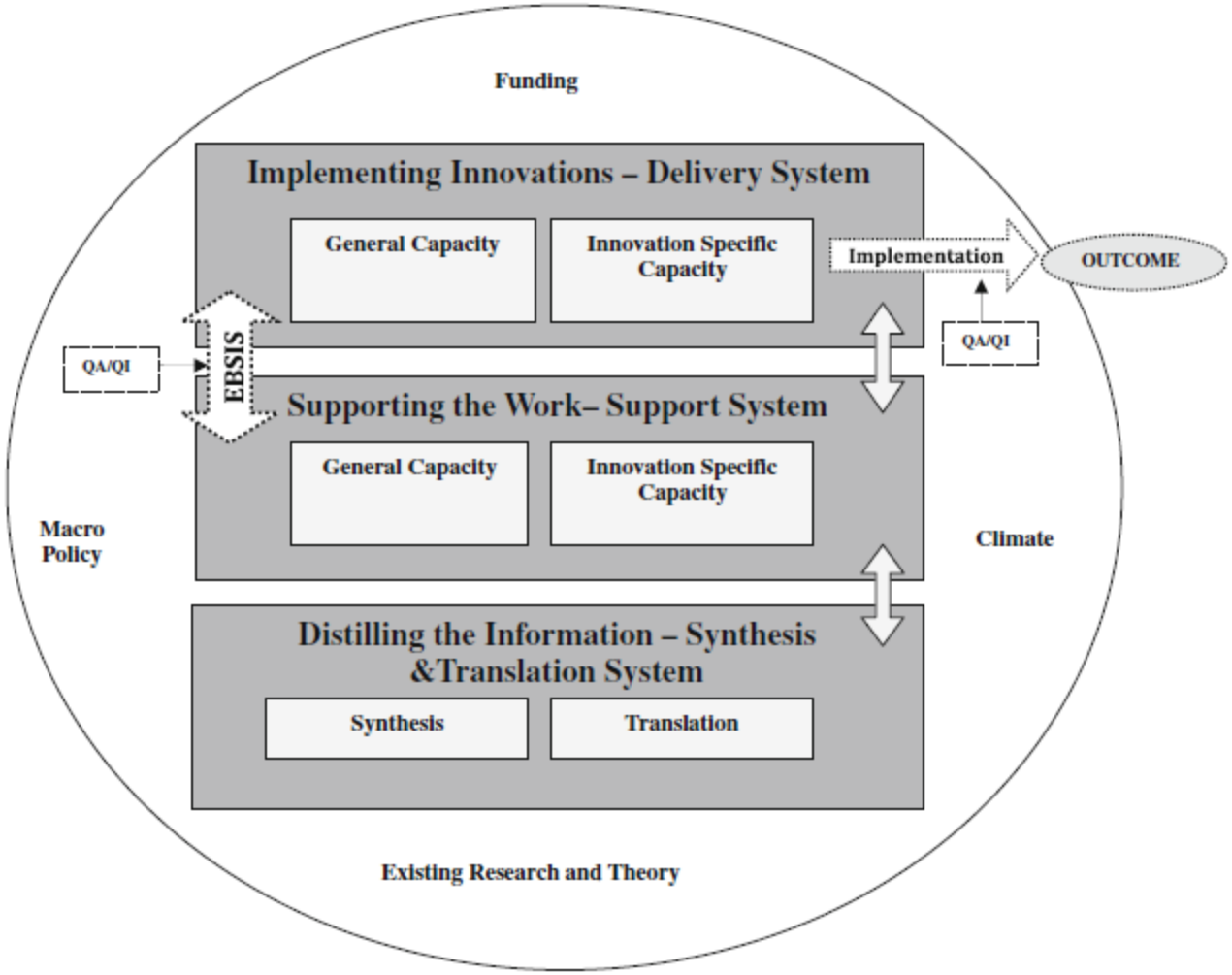
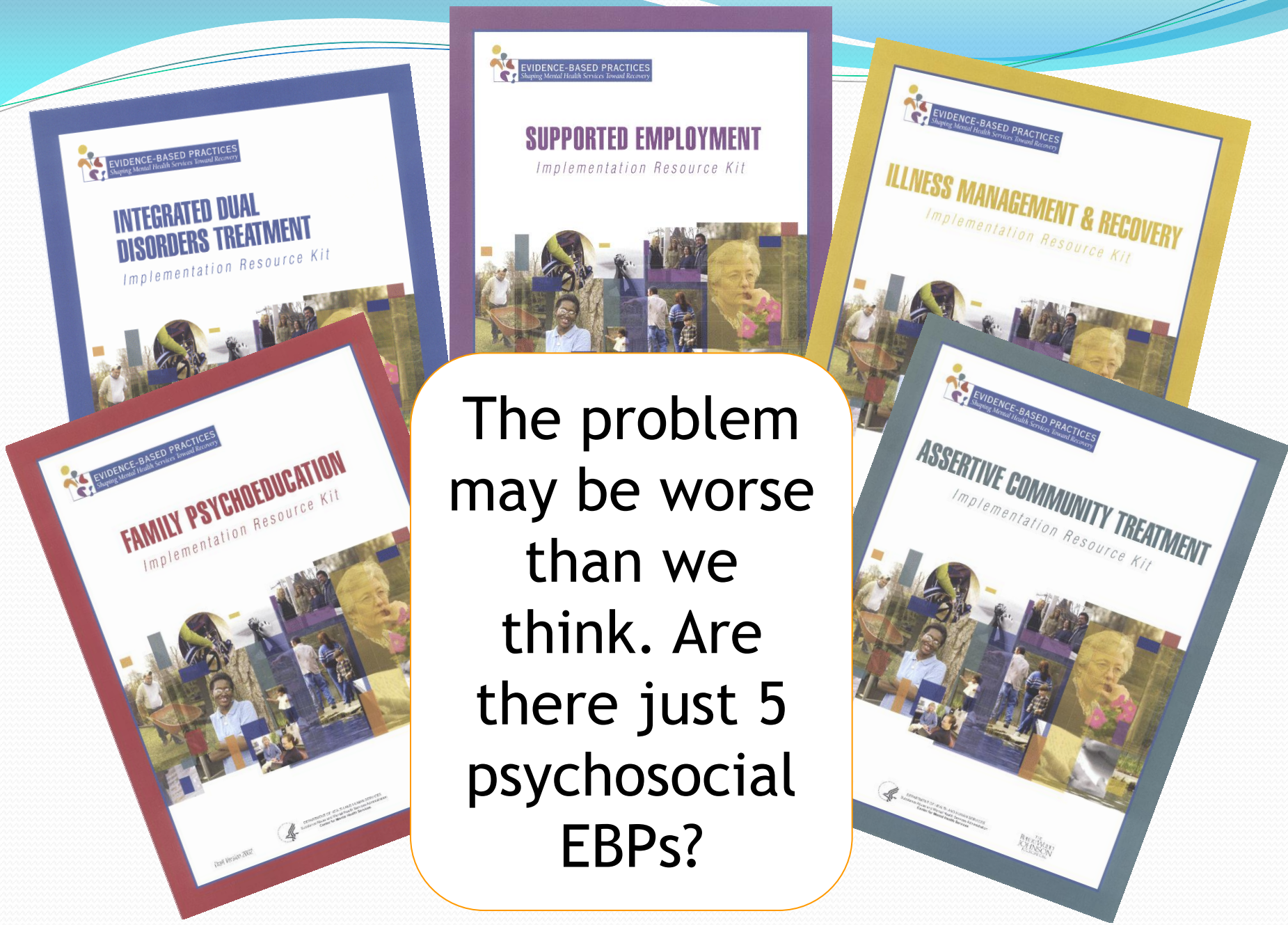


Fig. 1 Relationship between the EBSIS and the ISF. *Solid lines* indicate the original ISF (2008) figure and *dashed lines* indicate additions by our EBSIS approach. QA/QI are emphasized in two places: the provision of support to the Delivery System *and* the implementation of innovations (programs, policies, etc.)

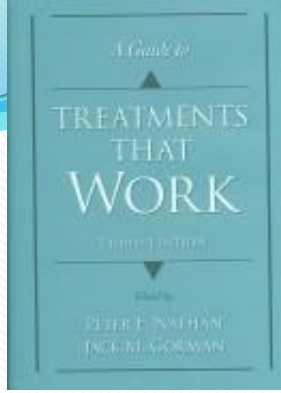
Background—the good news: Explosion of interest in EBPs

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

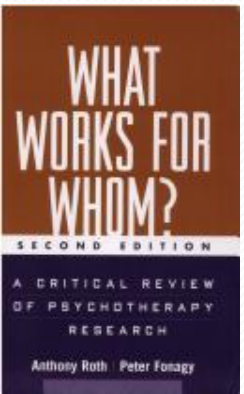


The problem may be worse than we think. Are there just 5 psychosocial EBPs?



Or, are there over 100?

Date	Review source	Number of EBPs
1995	Division 12 Taskforce	22 effective, 7 probable
1998	Treatments that Work	44 effective, 20 probable
2001	National EBP Project	6 effective
2001	Chambless, Annual Review of Psychology Article	108 effective or probable for adults; 37 for children
2005	What works for whom	31 effective, 28 probable
2007	Treatments that Work	69 effective, 73 probable
2008	SAMHSA Registry	38 w/ experimental support; 58 legacy programs



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)

*Evidence-based Practice Reporting for Uniform Reporting Service and National Outcome Measures Conference, Bethesda, Sept, 2007

Fidelity Assessment Variables

Mode	Face-to-face, Phone, Self-report
Data collection site	On-site Off-site
Data collector	External—outside assessor Agency affiliated—within agency, but outside the team Internal—self assessment by team/program
Instrument	Full/ partial/ screen
Data source	EMR, chart review, self-report, observation
Informants	Team leader, full team, specific specialties (e.g., nurse), clients, significant others
Team variables	Size, location, years of operation, developmental status

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)

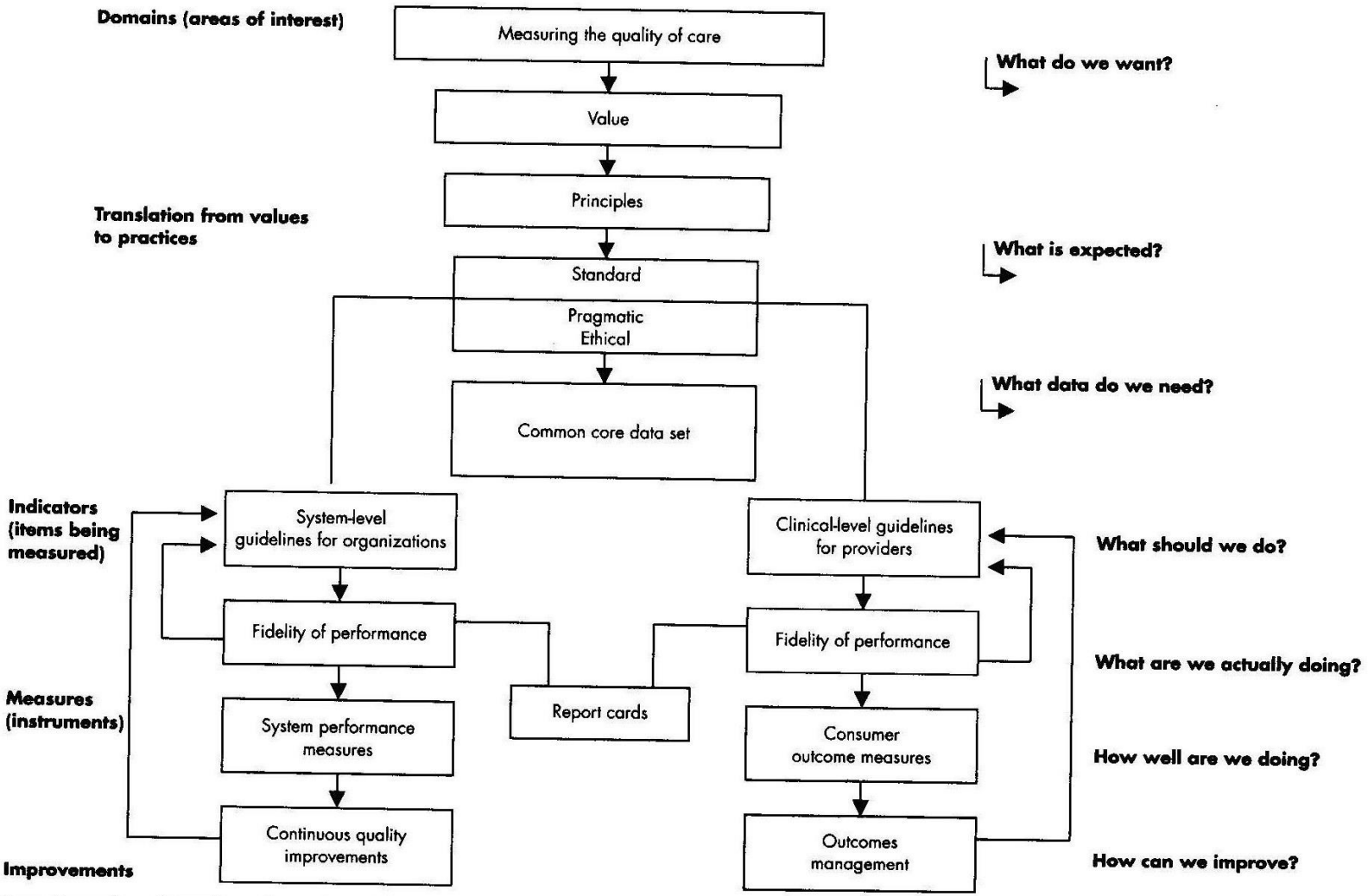
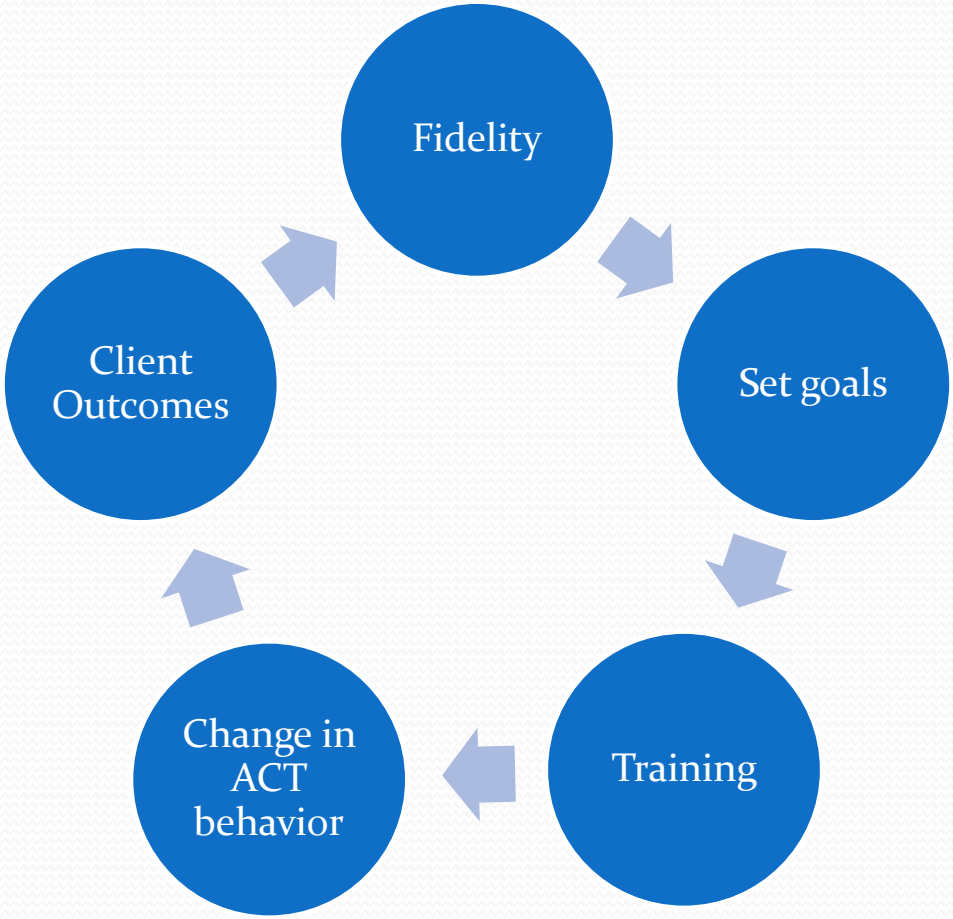
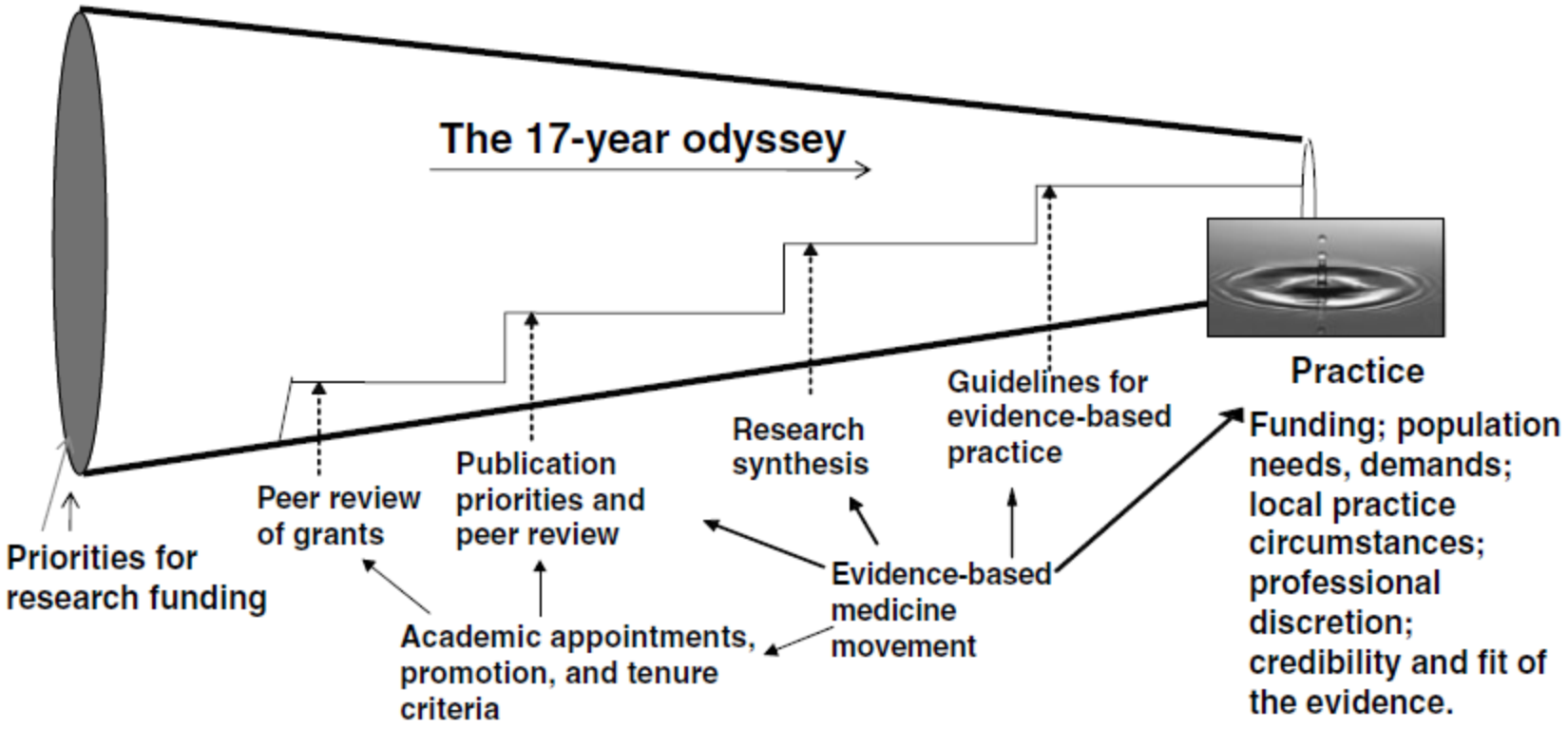


Figure 12–1. Quality process.

Manderscheid et al. (2001). Status of national efforts to improve accountability. In B. Dickey & L Sederer (eds) Improving mental health care: Commitment to quality. American Psychiatric Publishing, Washington DC

Future: Fidelity Outcome Training Model




The 17-year odyssey
Practice
Guidelines for evidence-based practice
Research synthesis
Publication priorities and peer review
Peer review of grants
Priorities for research funding
Funding; population needs, demands; local practice circumstances; professional discretion; credibility and fit of the evidence.
Evidence-based medicine movement
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?)

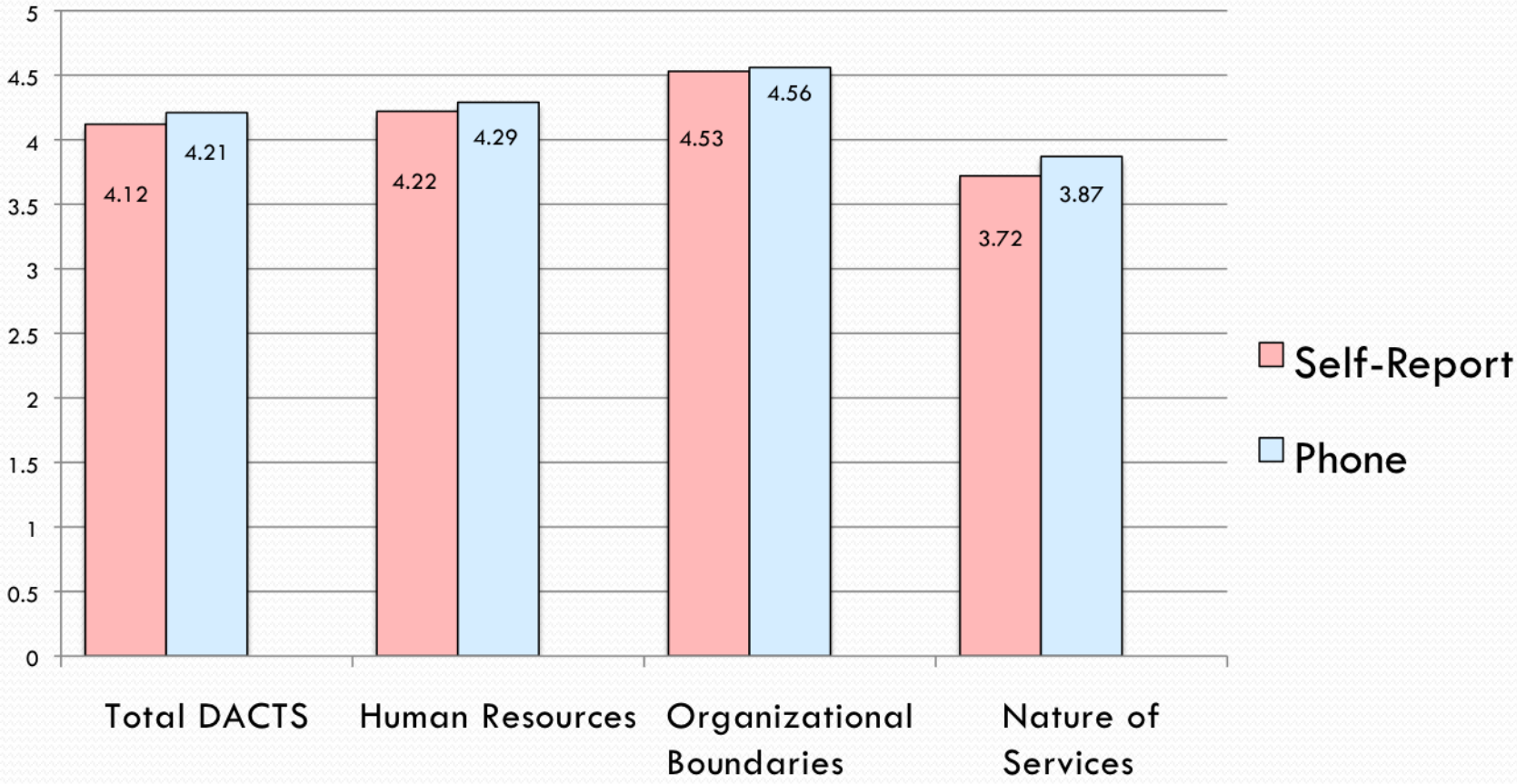
Assessment – Continuous rating

- 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

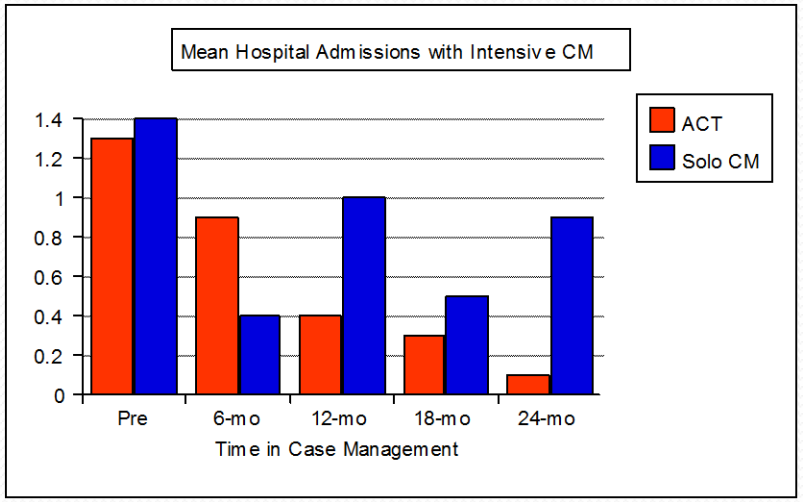
- Inter-rater reliability
  - Total and subscale scores for each rater
    - Intraclass Correlation Coefficient (ICC) between two raters of each fidelity method (consistency)
    - Mean and range of absolute value of differences between raters for each method (consensus)
- Validity
  - Total and subscale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



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



1.. Bond, G. R., Pensec, M., Dietzen, L., McCafferty, D., Giemza, R., & Sipple, H. W. (1991). Intensive case management for frequent users of psychiatric hospitals in a large city: A comparison of team and individual caseloads. Psychosocial Rehabilitation Journal, 15(1), 90-98.
2. Jerrell, J.M., & Ridgely, M.S. (1999). Impact of robustness of program implementation on outcomes of clients in dual diagnosis programs. *Psychiatric Services, 50*, 109–112.
3. Solomon, P., & Draine, J. (2001). The state of knowledge of the effectiveness of consumer provided services. Psychiatric Rehabilitation Journal, 25, 20-27.

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