HHS Public Access

Author manuscript

Adm Policy Ment Health. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Adm Policy Ment Health. 2022 July ; 49(4): 670–693. doi:10.1007/s10488-022-01191-5.

Performance of a Supervisor Observational Coding System and an Audit and Feedback Intervention

Jason E. Chapman[1], **Sonja K. Schoenwald**[1], **Ashli J. Sheidow**[1], **Phillippe B. Cunningham**[2]

[1]Oregon Social Learning Center

[2]Medical University of South Carolina

Abstract

Purpose.—Workplace-based clinical supervision is common in community based mental health care for youth and families and could be a leveraged to scale and improve the implementation of evidence-based treatment (EBTs). Accurate methods are needed to measure, monitor, and support supervisor performance with limited disruption to workflow. Audit and Feedback (A&F) interventions may offer some promise in this regard.

Method.—The study—a randomized controlled trial with 60 clinical supervisors measured longitudinally for seven months—had two parts: (1) psychometric evaluation of an observational coding system for measuring adherence and competence of EBT supervision and (2) evaluation of an experimental Supervisor Audit and Feedback (SAF) intervention on outcomes of supervisor adherence and competence. All supervisors recorded and uploaded weekly supervision sessions for seven months, and those in the experimental condition were provided a single, monthly web-based feedback report. Psychometric performance was evaluated using measurement models based in Item Response Theory (IRT), and the effect of the SAF intervention was evaluated using mixed-effects regression models.

Results.—The observational instrument performed well across psychometric indicators of dimensionality, rating scale functionality, and item fit; however, coder reliability was lower for competence than for adherence. Statistically significant A&F effects were largely in the expected directions and consistent with hypotheses.

Conclusions.—The observational coding system performed well, and a monthly electronic feedback report showed promise in maintaining or improving community-based clinical supervisors' adherence and, to a lesser extent, competence. Limitations discussed include unknown generalizability to the supervision of other EBTs.

Keywords

Clinical supervision; observational measurement; Audit and Feedback; mental health; behavioral health; evidence-based treatment; implementation

Correspondence regarding this article should be addressed to: sonja.schoenwald@oslc.org .

All research procedures were fully consistent with ethical guidelines and approved by the pertinent Institutional Review Boards.

Mental health services and implementation research have illuminated leverage points in routine care to significantly improve treatment outcomes. Among these are support structures such as clinical supervision and computerized information systems. Clinical supervision is a commonly used quality assurance method in community mental health care settings serving children (Bickman, 2000; Schoenwald et al., 2008), and it has been identified as a potentially potent and sustainable implementation strategy to support clinician delivery of evidence-based treatment (Chorpita & Regan, 2009; Dorsey et al., 2018). The measurement of effective supervision practices, however, is nascent, having occurred primarily in the context of treatment efficacy or effectiveness trials in which participating supervisors and clinicians may not reflect a community-based workforce. For example, in a multi-site effectiveness trial of several evidence-based practices for children, seasoned clinicians in community practice settings participated in supervision that was conducted by post-doctoral fellows with expertise in the practices (Bearman et al.,2013). In that study, active supervision techniques such as modeling and role play, when compared with case discussion alone, predicted greater clinician use of evidence-based practices in subsequent treatment sessions. Similarly, use of active learning techniques in Cognitive Behavioral Therapy (CBT) supervision has been associated with treatment fidelity among trainees participating in an analogue experiment (Bearman et al., 2017). Internationally, owing in part to government initiatives to increase the reach of effective treatments, efforts are underway to document and evaluate CBT supervision in routine care and identify contextual factors that enable such supervision (Newman et al., 2016).

In the U.S., observational studies have begun to illuminate the nature of routine supervision and extent to which it resembles the supervision in treatment trials,which is characterized by rigorous model-specific supervision and fidelity monitoring (Roth et al., 2010). Studies have indicated that clinical supervision is among effective methods to train clinicians to deliver evidence-based practices (Beidas & Kendall, 2010). In a recent study of routine supervision in two community mental health clinics, which focused on supervisor employees and early career clinician trainees, observational data revealed variable use of evidence-based supervision micro skills and limited competence in their delivery (Bailin et al., 2018).

In the first objective examination of workplace-based clinical supervision of EBT following clinician and supervisor EBT training, Dorsey et al. (2018) found (using observational coding) that supervision techniques typical of treatment trials were used rarely or with low intensity and that supervision content varied considerably, was tailored to individual clinicians, and was driven to some degree by individual supervisors. In an ongoing randomized trial, the investigators are evaluating the effect of two different supervision packages, each of which includes elements from treatment trials, on clinician fidelity and child outcomes (Dorsey et al., 2013). Their study is among few (see Bradshaw et al., 2007; Schoenwald et al., 2009) to have evaluated supervision effects on both clinician EBT delivery and client outcomes, and it is the most rigorous evaluation to date.

Computerized information systems are a common feature of mental health systems that can be leveraged to support the ongoing measurement, reporting, and improvement of treatment progress and outcomes. Specifically, considerable research supports the premise that measurement feedback systems (MFS) improve client outcomes (Bickman, 2020).

MFS are designed to measure, monitor, and feed-back information on client progress in psychotherapy. In meta-analyses, MFS have demonstrated small to moderate effectiveness in improving client outcomes overall, and particularly potent effects in averting deterioration among clients predicted to have poor outcomes (Lambert et al., 2018). They have been characterized as “digital interventions” that collect and provide clinically useful information to therapists in close to real time (Lyon et al., 2016). In mental health services organizations that adopt MFS, however, implementation is often hampered by logistical, professional, and organizational challenges, and clinician use is quite limited (Gleacher et al., 2016; Lambert & Harmon, 2018; Sale et al., 2021). Clinical supervision has been identified as a possible source of support for clinician use of MFS and associated client benefit (Lewis et al., 2019).

Outside of mental health systems, one of the most common and well-studied strategies to implement evidence-based health practices is Audit and Feedback (A&F) (Colquhoun et al., 2021). A&F interventions involve summarizing data about specific aspects of practice over a specified period of time and feeding it back to practitioners. There is substantial variability in the design, content, and delivery of A&F interventions in health care. A Cochrane review of 140 trials with health care professionals found A&F to show modest effectiveness across a wide range of applications and settings, although effect sizes varied considerably (Ivers, et al., 2012). The “who, what, when, why, and how much” of A&F interventions, along with their frequency of use in research trials, has since been catalogued (Colquhoun et al., 2017). The focal practices included, for example, prescription rates, vaccination rates, identification of patients with stroke, and ordering of unnecessary tests.

The Cochrane review identified conditions under which A&F may be most effective, including when: health professionals are not performing well at the outset, the person responsible for providing feedback is a supervisor or colleague, feedback is provided more than once, feedback is given both verbally and in writing, and feedback includes clear targets and an action plan A subsequent systematic review and cumulative analysis re-affirmed the relevance of these features to A&F effect sizes, as well as the paucity of these features in A&F interventions evaluated in 32 trials published since the Cochrane review (Ivers et al., 2014). Despite cumulative evidence that repeated feedback, feedback delivered by a supervisor or respected colleague, and feedback with explicit goals and action plans were associated with increased effect sizes, the authors noted that most of the 32 studies involved a single feedback cycle, only six involved supervisor- or colleague-delivered feedback, and none provided the explicit goals and action plans to meet them. The authors posit the potency of A&F interventions is likely to remain modest unless they include these features.

A&F interventions also hold promise for mental health care. For example, a recent study evaluated a statewide initiative to improve antipsychotic prescribing practices among psychiatrists in community mental health centers. The initiative included low-intensity academic detailing (i.e., a sequence of four, 50-minute in-person educational sessions delivered during regular administrative meetings over two years by experienced psychiatrists), with audit-and feedback on observed rates of polypharmacy and prescribing of medications. The observed rates appeared alongside information about recommended treatment, and practice improvements followed (Brunette et al., 2018). Because supervision is a standard feature of community mental health practice, it provides a natural opportunity

for A&F interventions. This has great potential for efficiency, as each supervisor affects multiple clinicians or case workers who, in turn, affect multiple clients or families. Supervisor-focused A&F was supported by a pilot study evaluating the R^3 supervision strategy (Saldana et al., 2016). In this study, supervisors in an urban child welfare system received once monthly feedback based on video-recorded group supervision sessions with caseworkers, and the results indicated that supervisor fidelity to the R^3 model improved over time.

The Role of Fidelity Measurement in Audit-and-Feedback

In the R^3 study, the purpose of the supervision strategy was articulated collaboratively by the service system and strategy developers, the video recording method used to observe supervision had been successfully deployed in prior studies of the community-based implementation of evidence-based interventions, and the method used to measure fidelity was designed to be relevant to and used by the supervisors and leaders in the system. These features of the fidelity measurement approach align with those ascribed to “pragmatic measures” in implementation research, in that they are “relevant to practice or policy stakeholders and feasible to use in real-world settings” (Glasgow& Riley, 2013). The need for measurement methods that are both effective (scientifically validated) and efficient (feasible and useful in routine care) (Schoenwald et al., 2011) is particularly apparent with respect to treatment fidelity. Treatment fidelity, also called treatment integrity (Southam-Gerow et al., 2021), is the extent to which a treatment was delivered as intended (Hogue et al., 1996). In psychotherapy research, the construct encompasses three components: adherence, the extent to which treatments as delivered include prescribed components and omit proscribed ones (Yeaton & Sechrest, 1981); competence, the skill with which treatment is delivered; and differentiation, the extent to which a treatment can be distinguished from others.

Despite its centrality to assessing the use of a particular treatment approach (or component, or technique), and thus to parsing success or failure of the treatment from success or failure of its application, adequate measurement of all three components of fidelity has occurred with low frequency in psychotherapy efficacy and effectiveness studies (Perepletchikova et al., 2007). A review of over 300 published studies in which fidelity or adherence assessment was mentioned found low rates of reporting psychometric properties (35%) and associations between adherence and outcomes (10%; and Schoenwald & Garland, 2012). A recent meta-analysis of 29 child treatment studies found a small but significant relationship between adherence and clinical outcomes, but no relationship between competence or composite integrity measures in the small subset of studies that included them (Collyer et al., 2020).

When considering the “voltage drop” in effectiveness for some EBTs deployed in community practice settings, loss of fidelity is a key factor (Chambers et al., 2013). There is evidence that fidelity to key components of EBT erodes quickly after a formal training (Farmer et al., 2016; Stirman et al., 2013). Additionally, fidelity has been identified as an implementation outcome in circumstances where it is clearly defined, at levels consistent with a valid standard, and associated with program outcomes (Landsverk et al., 2012). This suggests research is needed on effective strategies to support the fidelity of EBT

implementation in routine care (Stirman et al., 2012; Weisz et al., 2014). Observational data obtained during treatment trials suggest the provision of fidelity-focused feedback to clinicians holds some promise in this regard (Boxmeyer et al., 2008; Caron & Dozier, 2019; Lochman et al., 2009). In a study currently underway, the effects of two implementation support strategies on treatment fidelity and outcomes are being compared: fidelity-oriented consultation and continuous quality improvement learning collaboratives (Stirman et al., 2017). Accurate and low burden methods to measure fidelity are needed to generate such feedback and to support the larger scale implementation of effective psychosocial treatments. Federal research funding has been made available to support the development and evaluation of efficient and scalable methods to measure clinician fidelity in community practice settings (Beidas et al. 2016; Stirman et al., 2018). The current study directed fidelity measurement development and evaluation, and feedback efforts on the leverage point of clinical supervision.

The selected EBT was Multisystemic Therapy® (MST; Henggeler et al., 2009). MST is an evidence-based, intensive family-and community-based treatment originally developed for families of delinquent youth at imminent risk of incarceration or placement in other restrictive settings. MST is implemented nationally and internationally using the MST Quality Assurance/Quality Improvement (MST QA/QI) system, features of which have been detailed in prior publications (see, e.g., Henggeler et al., 2009; Schoenwald, 2016) and include training of therapists and supervisors, weekly group supervision, monthly collection of family-reported therapist adherence data, and bi-monthly collection of therapist-reported supervisor adherence data. Higher therapist adherence was found in randomized effectiveness trials to predict better long-term criminal and out-of-home placement outcomes and improvements in youth behavior and family functioning (Henggeler et al., 1997; Henggeler et al, 1999; Huey et al., 2000; Ogden & Hagen, 2006; Timmons-Mitchell, Bender, Kishna, & Mitchell, 2006). The relationship of therapist adherence to youth outcomes held in a prospective 43-site transportability study, in which statistically and clinically significant associations were also found among clinical supervision, therapist adherence, and youth outcomes (Schoenwald et al., 2009).

The MST QA/QI system is deployed by purveyor organizations (Fixsen et al., 2005). The purveyor organizations are MST Services, LLC (MSTS), which is licensed by the Medical University of South Carolina (MUSC) to disseminate MST technology, and its domestic and international Network Partners (NPs), the latter of which serve the majority of MST programs nationally and internationally. The scaling and sustainment of MST programs and evidence of relations among supervisor adherence measured indirectly (therapist reports every two months), therapist adherence, and youth outcomes laid the practical and empirical groundwork to develop and evaluate an observational method to measure evidence-based supervision deployed in community practice settings and preliminary effects of and audit-and-feedback system on supervisor performance.

Method

In the present study, we conducted a detailed psychometric evaluation of an observational coding system to measure supervisor fidelity to an EBT supervision protocol in community-

based settings that was developed and pilot tested in a prior, related study; and evaluated the effect of an experimental supervisor A&F intervention on supervisor adherence and competence as measured using the observational coding system. This prospective, two-arm Randomized Controlled Trial (RCT) began in September of 2014 and supervision sessions were recorded and uploaded through August of 2016. All study procedures were approved by the Institutional Review Board of the Oregon Social Learning Center.

Recruitment Procedures

Eligible participants were MST supervisors employed in domestic MST programs who had not participated in the initial Supervisor Observational Coding System (SOCS) measurement development study (see Appendix A). Supervisors were identified with permission from, and in collaboration with, purveyor organizations, which provided supervisor names and contact information. From this information, supervisors were randomly selected to be approached for study recruitment. Because multiple supervisors could receive consultation from the same MST trainer, the goal was to recruit no more than three supervisors for a given trainer. This was mostly successful, with only two instances of trainers with four supervisors recruited. The remaining trainers had one to three supervisors recruited.

A study investigator contacted supervisors by email to describe the study and inquire about the supervisor's interest in participating. For supervisors expressing interest, the project coordinator scheduled a telephone call to provide more information about the study, answer supervisor questions, and initiate the informed consent process. Prior to the scheduled informed consent call, the project coordinator emailed a PDF of the consent form to each supervisor and mailed a hard copy of the form with an enclosed self-addressed envelope. Upon completion of the call, the supervisor mailed the signed form to the project coordinator, who signed and mailed a copy of the fully executed form back to the supervisor, retaining the original. When a supervisor declined to participate, the next eligible participant was randomly selected from the list. As detailed in the CONSORT diagram (Figure 1), of 83 supervisors contacted, 60 (72%) provided informed consented to participate. Among the 23 supervisors who declined to participate, eight declined during the informed consent call, eight decided not to schedule a consent call after receiving the form, and seven did not respond to requests for a call after receiving the form. Reasons individuals declined to participate included imminent maternity or other health related leave; retirement or promotion to a position other than supervisor; lack of approval from the provider organization executive director, board, or compliance department; or supervisor concerns about the logistics and potential burden of study procedures.

Participants

Sixty MST supervisors were recruited, and of these, three (5%) did not provide any data (two left their position prior to the start of recording and one withdrew due to their supervisor not having approved participation). The remaining 57 were employed by 49 provider organizations. Some supervisors elected not to respond to all demographic questions. Of those responding, most identified as female (86%; 0 missing), with an average age of 36.5 years (SD = 6.42; range 29 – 56 years; 0 missing). With respect to ethnicity, 75% identified as *not* Hispanic or Latinx, 8% as Hispanic or Latinx, with the remainder

being unknown. For race, 55% identified as White, 28% as Black or African American, 3% as Asian or more than one race, with the remainder being unknown. All held a master's degree (1 missing), and 96% (4 missing) were employed full-time, with the remaining 4% employed part-time. Supervisors' average annual salary was $49,900 (*SD* = $12,898; 7 missing). With respect to current and past employment in MST programs, 47.2% (4 missing) had been employed by the current program for more than four years, 15% for 3–4 years, 11% for 2–3 years, and 13% for 0–1 year, respectively. In addition, 53% of the supervisors had been employed in an MST program (even if not the current one) for over four years. Over half (57%; 3 missing) were employed by mental health services organizations, 15% by social services organizations, 15% by other or multi-purpose organizations, 11% by juvenile justice organizations, and 12% by substance abuse services organizations.

Intervention Conditions

Each supervisor was randomly assigned to one of two conditions: Supervisor Audit (SA) or Supervisor Audit and Feedback (SAF). The randomization procedure used a random number generator (SPSS, Mersenne Twister algorithm with random starting values) to assign 60 numeric IDs to the SA or SAF condition (30 per condition). The IDs were then assigned to supervisors based on the order of consent to participate in the study. In both the SA and SAF conditions, supervisors recorded and uploaded weekly group supervision sessions for seven consecutive months. Each supervisor was provided with a digital recorder purchased with research grant funds, along with instructions for operating the recorder and uploading recordings to the study's secure website. The project coordinator also provided individual phone-based training on use of the recorder and the upload process.

Supervisors in the SA condition uploaded session recordings and received no further information about their supervision. In the SAF condition, supervisors received monthly feedback for six consecutive months about their Adherence and Competence in a session that was randomly selected for coding each month. Specifically, the project coordinator sent an email to notify SAF supervisors that feedback was available, along with instructions for accessing the web-based report. Consistent with design principles underlying technological advances in measurement-based care and clinical decision support (Ivers et al., 2014; Landis-Lewis et al., 2020), the report was designed to profile case discussion performance in an easily interpretable way that reflected priorities for future sessions. With minimal use of text, the report primarily employed graphics and color-coding to summarize and convey information. At the top of the form, a Timeline depicted the dates of coded sessions, along with a link to an electronic interpretation guide. Next, for the focal session, a Session Highlights section provided examples of components with Competence ratings (if applicable) of Low (Red), Moderate (Yellow) or High (Green), along with components that were rarely or never observed. The final three sections summarized the components coded in each of the three supervision domains (detailed in the Supervisor Observational Coding System section). In each domain, the components had a brief label (with a full definition available via a clickable *information* icon). For each, a color-coded graphic illustrated the number of cases where the component was not observed or, when observed, the number where Competence was Low (Red), Moderate (Yellow), and High (Green). For each report, examples were specified by the first author and based on reviewing the

supervisor's performance in the coded session, along with feedback from prior reports. This required approximately 15 minutes per report, and it was estimated that reviewing the report would require 15 minutes for the supervisor. An example feedback report is provided in Appendix B.

Instruments

Personnel Data Inventory—The Personnel Data Inventory (PDI; Schoenwald, 1998) captures demographic, educational, and professional experience data from therapists and supervisors in MST programs. The project coordinator mailed the PDI to participants with a self-addressed stamped envelope and requested return date that preceded the start of the digital recording and upload procedures.

Supervisor Observational Coding System—The Supervisor Observational Coding System (SOCS) was developed in the first study (R21 MH097000; J. Chapman & S. Schoenwald, Co-PIs) using IRT-based development and evaluation methods. The measurement development process, including revisions prior to the present study, are detailed in Appendix A. The psychometric functioning of the SOCS is a primary focus of the present study, and the scores for analysis are described in the Data Analysis Strategy.

Domains and Items.: The SOCS components are presented in Table A1. The instrument is comprised of three theoretical domains: Analytic Process (AP) with 10 components, Principles (P) with 9 components, and Structure and Process (SP) with 12 components. All AP and P components received two ratings. The first was for Adherence, which indicated whether the component was delivered during the observation period (0 = No, 1 = Yes). The second was for Competence which, for components that were delivered, reflected the quality of that delivery (1 = Low, 2 = Moderate, 3 = High). For SP, the components were always applicable (with only two exceptions), and as such, they were only rated for Competence.

Rating Structure.: The group supervision sessions were structured with a series of individual case discussions between the supervisor and each therapist on the team. As such, the AP, P, and SP domains were rated for each of the first six case discussions in the session. This resulted in six sets of ratings for most sessions (81.8%), with the remainder being comprised of five (7.5%), four (6.4%), or one-to-three (4.3%) case discussions, respectively. The average duration of a single case discussion was 8.7 minutes (SD = 6.1), and the average duration of the rated cases was 49.2 minutes (SD = 18.6). The approximate time requirement for coding was the duration of rated cases plus 10 minutes for pausing the recording, consulting the coding manual, reviewing and finalizing codes, and/or adding comments. For sessions rated by more than one coder, the start and end times for each case were reviewed to confirm that the same six case discussions were rated, and any discrepancies were resolved and re-rated prior to data entry.

Coders and Coder Training.: There were four coders, all of whom had a master's degree in counseling or clinical psychology. Training included assignment and review of existing materials on the MST supervision protocol, review and discussion of the SOCS coding manual, review of the coding manual in conjunction with exemplar audio segments, group

coding of exemplar audio segments for each domain and component, individual coding of segments with group review and discussion, and coding of full sessions with group review and discussion. The focus was on achieving absolute agreement across coders relative to existing ratings from expert coders. The process continued until the ratings from independent coding of the same session yielded acceptable levels of agreement across coders and saturation on factors leading to disagreements in ratings. The training protocol required approximately 60 hours per coder. As detailed subsequently, to combat challenges related to absolute agreement for Competence ratings, the coding plan was revised to include a substantial proportion of sessions assigned to two coders.

Coding Plan.: Supervisors uploaded audio recordings of group supervision sessions on a weekly basis for seven months. From these, one session was randomly selected per supervisor per month for the purpose of observational coding. Study condition was masked for all assignments. Sessions were assigned via email on a weekly basis, with a one-week window for completion. Coders received an average of 5.5 assignments per week (SD = 2.1, Min. = 2, Max. = 14). To complete coding assignments, coders logged in to the secure project website, navigated to the assigned supervisor and session, streamed the recording, and recorded ratings on a paper form. The coding plan ensured that each coder rated an approximately equal number of sessions for each supervisor and that all pairs of coders were evenly balanced across supervisors and study months. Sessions were assigned by rating domain, with AP and SP linked and P as a separate domain. This strategy was largely to manage the coding burden for each assigned session. Within these criteria, the individual sessions were assigned randomly to coders. Informed by the pilot study (see Appendix A), some challenges were expected with respect to interrater agreement on Competence ratings. Because of this, the percentage of sessions assigned to a second coder was increased, with a target of three sessions per supervisor. This resulted in three or four IRR sessions for 49 of 57 supervisors (85.9%), with the remainder having one or two IRR sessions.

Uploaded and Coded Sessions.: Fifty-seven supervisors (95%) uploaded at least one session recording. A total of 1,039 *weekly* audio recordings were uploaded, representing 83% of all eligible weekly sessions across supervisors. From these weekly sessions, one per supervisor was randomly selected *each month* for the purpose of observational coding. Across supervisors, 97% of eligible months had at least one uploaded session for coding. Complete data were available for 86% of supervisors (n = 49), with 12% of supervisors (n = 7) providing 71%-86% of the expected data and the remaining 2% (n = 1) providing 57% of the expected data. In the final data, 79% (n = 45) of supervisors had coded sessions for seven months, 10.5% (n = 6) for six months, and 10.5% (n = 6) for three to five months.

Data Analysis Strategy

Data Structure—The research design and coding system led to a data structure that was extensively nested, though for analysis, the specific structure varied somewhat from model to model. Generally, there were six case discussions (level-1; mean = 5.7, SD = 0.5; C_{ases} = 2,110) nested within seven supervision sessions (level-2; mean = 6.6, SD = 1.0; $S_{essions}$ = 374) nested within 57 supervisors (level-3). There was also cross-classification by coder.

Each model reported in the Results section includes a description of the relevant data structure for analysis.

Measurement Models—The psychometric performance of the observational coding system was evaluated using a series of measurement models based in Item Response Theory, including Rasch models (Wright & Mok, 2000; Rasch, 1980), Many-Facet Rasch Models (MFRM; Linacre, 1994), and mixed-effects formulations of Rasch-equivalent models known as hierarchical generalized linear measurement models (HGLMMs; Kamata, 2001). The models were implemented using multiple software packages: WINSTEPS (Linacre, 2019b), FACETS (Linacre, 2019a), HLM (Raudenbush et al., 2013), and MLwiN (Charlton et al., 2019). The Rasch (or Rasch-equivalent) model is a probabilistic measurement model where, using Adherence as an example, the probability of a specific *supervisor* delivering a specific supervision *component* is the net result of the supervisor's overall level of Adherence (i.e., "ability") and the component's overall rate of delivery (i.e., "difficulty"). Thus, for a supervisor with low Adherence and a component that is rarely delivered, the probability of delivery would be *low*. Likewise, for a supervisor with high Adherence and a component that is frequently delivered, the probability of delivery would be *high*. The standard model is straightforwardly extended to accommodate other data features, including rating scales with three or more categories, additional model "facets" (e.g., cases, sessions, coders), and nested data structures (e.g., sessions within supervisors). The model results provide multiple indicators of psychometric performance, including: dimensionality, inter-rater agreement, rating scale performance, item fit, reliability and separation, and the degree to which the items are suitable for assessing the sample of observations. Supplementary models provided estimated variance components and multilevel reliability statistics. Each indicator is described in the Results. It is important to note that other measurement models, such as two-parameter logistic or graded response models, are available. However, Rasch models confer practical benefits for modest sample sizes and nested data, and the MFRM is specifically intended for evaluating rater data (e.g., Myford & Wolfe, 2003).

Adherence and Competence Scores—Two versions of Adherence and Competence scores were used to evaluate the experimental SAF intervention: logit-based Rasch measures and raw (average) scores. The decision to use both scores was informed by feasibility; specifically, although logit-based scores were ideal measurement-wise, any applications of the SOCS in routine care would likely rely on raw scores. Raw average scores were computed as the average response across coders for each case discussion. The logit-based Rasch measures were obtained from HGLMMs implemented in MLwiN software. Adherence ratings were structured with item responses (level-1) nested within case discussions (level-2) nested within supervision sessions (level-3) nested within supervisors (level-4). Competence ratings were structured similarly, but with three ordered categories, the model had two dichotomous responses (thresholds) nested within each item response. Each model included a series of dummy-coded indicators to differentiate each item which, combined with the binomial outcome distribution and logit link function, resulted in a nested, Rasch-equivalent measurement model. The resulting item parameters are equivalent to Rasch item difficulty estimates. The parameters for case discussions—which are the focal

outcomes—are provided by empirical Bayes residuals, computed as the sum of supervisor, session, and case discussion residuals (Kamata, 2001; Ravand, 2015).

Prediction Models—The outcomes were structured with scores for case discussions (level-1) nested within months (level-2) nested within supervisors (level-3). The nested data were addressed using mixed-effects regression models (Raudenbush & Bryk, 2002) implemented in HLM software (Raudenbush et al., 2013), and all scores were modeled according to a normal sampling distribution. With the modest sample of supervisors, significance tests for fixed effects used asymptotic standard errors to compute the Walt test statistic (i.e., β/SE; Maas & Hox, 2005). Random effects were specified based on the likelihood ratio test (Singer & Willett, 2003). To model change over time, a linear polynomial—the number of months from baseline—was entered at the level of repeated measurements. The SAF intervention effect was modeled at supervisor-level using a dummy-coded indicator (0 = SA; 1 = SAF), with a cross-level interaction specified between intervention condition and the time term. This formulation tested for a difference between SAF and SA in baseline scores and in the linear rate of change over time. Supplementary models used dummy-coded indicators for each month following baseline (rather than a polynomial term), which tested for between-group differences in the change from baseline to each subsequent month.

Results

Psychometric Performance of the SAF Observational Coding System

Dimensionality—A key assumption of IRT-based measurement models is that the data are effectively unidimensional. For the theoretical AP, P, and SP domains, dimensionality was evaluated based on a principal component analysis of standardized Rasch item-person residuals (Bond & Fox, 2015; Smith, 2002). If dimensionality is not meaningful, the residuals should reflect random noise. To assess this, the Rasch PCA attempts to identify structure within the residual matrix, as this would reflect the presence of dimensions beyond the primary dimension being measured. An eigenvalue < 2.0 for the first contrast indicates that the data are reasonably unidimensional (Linacre, 2019b). The analysis utilized case-level ratings that were treated as independent observations (i.e., the model did not address nesting). In each case, the eigenvalues indicated that the theoretical domains were effectively unidimensional, with AP Adherence = 1.3, AP Competence = 1.4, P Adherence = 1.3, P Competence = 1.4, and SP Competence = 1.7. The results that follow are based on separate analysis of each domain.

Inter-Rater Agreement—When using an observational coding system to measure Adherence and Competence, the goal is for coders to rate each observation in an identical manner. Thus, the coders' role is one of a "rating machine" rather than an "expert judge" (Linacre, 2019). Accordingly, the primary indicator of inter-rater agreement was particularly stringent: the rate of absolute agreement. For Adherence, absolute agreement was high, both for the AP (89.4%) and P (86.6%) domains. From a more conservative perspective—which excluded agreement on components that *did not occur*—agreement was somewhat lower for AP (73.1%) and P (67.4%). For Competence, absolute agreement could only be computed

for items on which both coders indicated occurrence, and the rates were lower for AP (53.5%), P (52.6%), and SP (63.0%).

Rating Scale Functioning—Rating scale functioning was evaluated based on the results of the MFRM, which included facets for items, cases, sessions, supervisors, and coders. For Adherence ratings, the percentage of components rated as being delivered in both the AP and P domains was 35%, and there was no indication of misfitting rating categories. For Competence ratings in the AP domain, the percentage of ratings as *Low*, *Moderate*, and *High* was 27%, 45%, and 28%; in the P domain, 25%, 48%, and 27%; and in the SP domain, 20%, 53%, and 28%. For Competence ratings, the rating scale analysis also tested whether each category was consistently interpreted across coders and well-differentiated from adjacent categories (Wright & Masters, 1982). Specifically, for a three-point scale, the middle category should be the most probable rating for a ≥1.4 logit span of the Competence construct (Linacre, 2002). This was the case in each domain, with *Moderate* being the most probable rating for 1.5 (AP), 1.8 (P), and 2.4 (SP) logits. Likewise, there was no evidence of coders interpreting categories in the wrong order or utilizing them in significantly unpredictable ways.

Item Fit—The MFRM provides multiple indices of item fit, and the present results focus on Infit and Outfit mean square statistics. Both Infit and Outfit identify items characterized by unpredictable responses, with the difference being the type of observations (i.e., cases, sessions, supervisors) for whom the item is misfitting. Infit captures unpredictable responses on items among observations that are well-targeted to the item (e.g., a component of average difficulty for case discussions with average Adherence). In contrast, Outfit is sensitive to unpredictable responses from observations with more extreme scores (e.g., a component of average difficulty for case discussions with low Adherence). Infit and Outfit were evaluated relative to a threshold of 1.4 (Linacre & Wright, 1994). For Adherence and Competence, no items exceeded the threshold for the AP or P domains. For Competence ratings on the SP domain, three items were similarly misfitting on both Infit and Outfit: SP01 (1.79, 1.78), SP02 (1.63, 1.62), and SP03 (1.44, 1.52). These items were retained for computing scores because they index essential components of supervision that are prescribed to occur early in the supervision session (agenda setting, updates, top clinical concerns). This makes them distinct from components that occur throughout the session, which is the likely source of misfit.

Reliability and Separation—Rasch measurement models provide a range of reliability estimates. The first two types of estimates are highly conservative because they reflect the reliability for individual case discussions rather than supervision sessions. The first is Rasch separation reliability, which ranges from 0 to 1 and is interpreted consistently with traditional estimates of internal consistency. For Adherence, Rasch separation reliability for AP was .62 and for P was .55. For Competence, reliability was .52 for AP, .52 for P, and .48 for SP. The second indicator is the Rasch separation index, which is *not* bound by 1 and indicates the number of meaningful distinctions that can be made in the Adherence or Competence continuum based on the sample of items (e.g., one distinction would differentiate two levels of the construct). The ideal number depends on the intended use

of the instrument, and in the present case, 1–2 distinctions would be sufficient for informing supervisor feedback and evaluating the SAF intervention. For Adherence, the separation index was 1.3 for AP and 1.1 for P, and for Competence, it was 1.1 for AP, 1.0 for P, and 1.0 for SP. Thus, across the two indicators, reliability for an individual case discussion was modest, and the level of precision was suitable for distinguishing between two levels of Adherence or Competence.

Supplementary models were performed in HLM software to estimate the reliability of session- and supervisor-level Adherence and Competence scores (see Table 1), specifically, with Rasch measures or raw scores for case discussions (level-1) nested within sessions (level-2) nested within supervisors (level-3). These multilevel reliability estimates reflect both precision and variability (Singer & Willett, 2003). For Adherence, the reliability of session-level scores ranged from .27 to .76, and the reliability of supervisor-level scores ranged from .78 to .95. For Competence, session-level reliabilities ranged from .64 to .95, and supervisor-level reliabilities ranged from .48 to .79. In all cases, reliabilities were higher for Rasch measures relative to raw scores.

Effect of Supervisor Audit-and-Feedback Intervention on Supervisor Adherence and Competence

As detailed previously, two versions of Adherence and Competence session scores were evaluated as outcomes for the experimental SAF intervention: logit-based Rasch measures and raw average scores. The resulting data were structured with case discussions (level-1) nested within sessions (level-2) nested within supervisors (level-3) and analyzed using mixed-effects regression models. For each outcome, a preliminary unconditional model (i.e., random intercept only) was performed to estimate variance components and compute the proportion of variance attributable to each level of nesting. The estimates are reported in Table 1. Two types of prediction models were performed, one testing linear change over time (Table 2) and the other testing change between baseline and each later occasion (Table 3).

Analytic Process—In the linear growth models for AP Adherence (Table 2), SAF and SA did not differ significantly at baseline on either the Rasch or raw scores. However, on both scores, SAF and SA differed significantly on their linear change over time, with the rate of change for SAF being more positive than the rate of change for SA. Specifically, SA had a significant decrease in AP Adherence over time, but for SAF, Adherence did not change. When testing for change between baseline and each later month (Table 3), the pattern of findings was consistent. The difference in change for SAF and SA was statistically significant at months 2, 3, 5, and 6 (Rasch) and 2, 5, and 6 (raw). Specifically, SA had significant decreases at months 5 and 6 (Rasch) and months 2, 5, and 6 (raw), and SAF did not change significantly.

For AP Competence, SAF and SA did not differ significantly at baseline or on their linear rates of change over time (Table 2). When testing for change from baseline to each later month (Table 3), the groups did not differ significantly, but SAF had a significant increase in Competence at month 3 (Rasch). There were no significant effects based on raw scores.

Principles—For P Adherence, SAF and SA did not differ significantly at baseline or on their linear rates of change over time (Table 2). When testing for changes between baseline and later months (Table 3), the results were consistent for the Rasch measures and raw scores. Specifically, relative to SA, the change for SAF was significantly more positive from baseline to month 2. From baseline to month 4, the two groups did not differ, but SA had a significant decrease in P Adherence.

For P Competence, SAF and SA did not differ significantly at baseline or on their linear rates of change over time (Table 2). When testing for change from baseline to later months (Table 3), compared to SA, SAF had a significantly greater decrease in P Competence from baseline to month 2 (Rasch).

Structure and Process—For SP Competence based on Rasch measures, there was effectively no variability across case discussions within sessions. As such, the model was reduced to a two-level structure with sessions (level-1) nested within supervisors (level-2). Based on Rasch measures, SAF and SA did not differ significantly at baseline or on the rate of change over time, but for raw scores, SP Competence increased significantly more for SAF relative to SA. When testing for change from baseline to each later month, the groups did not differ at month 2, though SAF had a significant decrease. However, at month 3, the change for SAF was significantly more positive than the change for SA. When considering raw scores, SAF had significantly greater increases than SA from baseline to months 3, 5, and 6.

Discussion

This study evaluated an observational method to measure the performance of clinical supervisors with respect to Adherence and Competence and the effects of an A&F intervention on that performance. The measurement method, SOCS, was previously developed, evaluated, and revised in accordance with *Standards for Educational and Psychological Testing* (SEPT; AERA, APA, NCME, 1999) guidelines and associated IRT methods in a collaborative process that included treatment content experts, measurement experts, and individuals who provide training and ongoing consultation to clinical supervisors and therapists. The SOCS was evaluated across a range of psychometric indicators using Rasch-based measurement models that accounted for nested data and rater effects.

Indicators of the feasibility of use of the SOCS include the high proportion of sessions uploaded by supervisors over an extended time (seven months), and near equivalence of coding time and session length achieved by coders who had no prior experience with the supervision protocol, focal intervention, or observational coding. The experimental electronic feedback report profiled case discussion performance in an easily interpretable way that reflected priorities for future sessions. With minimal use of text, the report primarily employed graphics and a simple red-yellow-green color code to summarize and convey information. Supervisors received a single prompt regarding the availability of the report and could view it at their convenience.

The SOCS Measurement Model

The three theoretical dimensions—Analytic Process, Principles, and Structure and Process—were supported, with no indication of further dimensionality. Coders had a high rate of agreement when rating Adherence. However, for Competence, the three-point scale performed as intended, but coder agreement was lower. Across the AP and P domains, the items performed as intended. For SP, three items evidenced some degree of "misfit" to the model. The pattern of fit statistics suggests that these items assess content distinct from the main SP domain. Specifically, these items pertain to agenda setting and updates, which occurs early in the sessions and/or case discussions and could be delivered with a level of competence that was independent of overall SP Competence. For reliability, scores for individual *case discussions* were modestly reliable, but *session* scores (i.e., across case discussions) were more reliable. The reliability of supervisor-level scores was high, but this level of scoring would have fewer practical uses. In sum, the SOCS largely performed well and as intended, with the main challenge being coder agreement on Competence ratings. A strength of the results is that the underlying measurement models were sophisticated IRT-based models that accommodated key features of the data, which included multiple levels of nesting, multiple coders, a high rate of sessions that were double-coded, and ordered categorical rating scales.

The reliability of coder ratings for Competence was notably lower than it was for Adherence. This may have been due in part to the lower number of instances in which Competence could be rated; rater agreement on Competence could only be rated for items on which both coders indicated Adherence. The construct of competence has been more difficult than adherence to define and rate reliably in studies of therapist fidelity to EBT (Hogue et al., 2008; Perepletchikova et al., 2007). With respect to MST, whereas supervisor adherence has been reliably measured using indirect methods (therapist reports), the SOCS is the first instrument to attempt measurement of supervisor competence. Accordingly, we are circumspect with respect to the interpretation and implication of competence findings in our discussion of the RCT results.

SAF RCT Findings

In the context of an observational system designed to assess clinical supervision in community settings, providing supervisors with a single, monthly web-based feedback report led to some changes in supervisor Adherence and Competence. Statistically significant findings were, with one exception, in expected directions, and in most instances, IRT-based logit scores and raw average scores yielded similar patterns of findings.

For Adherence in the Analytic Process (AP) and Principles (P) domains, scores in the SA condition either remained steady or worsened over time, whereas for the SAF condition, scores remained steady or improved. This suggests that the feedback report may have reinforced supervisor use of prescribed components of supervision. Few effects on Competence were observed across the AP, P, or SP domains, and one was unexpected. In the SAF condition, Competence in the AP domain improved at month 3 relative to baseline, but in the Principles domain, it decreased at month 2. Combined, this may suggest the feedback report led supervisors to attempt to deliver more intervention components (i.e., increase in

Adherence), but delivered the components less well (i.e., decrease in Competence). Recall, however, that Competence can only be rated when components are delivered; this means that delivery of few components with high competence would lead to better Competence scores than delivery of more components with mixed competence. Competence findings in the Structure and Process (S&P) domain were consistent with expectations. In the SAF condition relative to SA, changes in raw scores between baseline and months 3,5, and 6 were more positive, as was the change in Rasch score change baseline to month 3.

The findings supporting SAF effects on supervisor Adherence and Competence suggest the feedback was sufficiently specific and timely to be actionable and potent. Given the focus of the feedback was a single, randomly selected supervision session that occurred within the month the report was provided, these findings suggest promise with respect to the potential efficiency and feasibility of its use. However, the efficiency and feasibility of the feedback report may also have contributed to the relatively limited number of significant effects as well as variability of effects over time. Positive changes favoring the SAF condition occurred at some—but not other— months relative to baseline. It is possible that a weekly feedback report—one that summarized Adherence and Competence performance over all four supervisions in the month—would have yielded more consistent effects. Such alternatives would require significantly more human and computational resources than did monthly feedback, not only for producing feedback but also for the supervisors receiving feedback. It is also possible that supervisors were selective about which feedback to heed.

Future use of the SAF system includes feasibility considerations, as the current version is not fully automated and there are other up-front and ongoing resource requirements. Specifically, for A&F, the system requires trained observational coders, a secure upload website, secure server space, routine data entry, programming of the feedback report, and ongoing specification of feedback content for each recipient. Some components, such as specifying feedback content, are well-positioned for automation, whereas others, such as training and supervision of observational coders, typically would require ongoing resources. For A&F recipients, resource considerations include the cost of access to the system, recording equipment, computer resources, and internet access.

Limitations

The study is characterized by several limitations. First, with a sample size of 60 supervisors, the RCT was powered to detect large effects, and it is possible additional significant findings would have emerged in a larger sample. Second, because this is the main test of the SAF, we used intent to treat analyses and did not evaluate effects of supervisors' actual use of feedback. Third, although we invited supervisor input in the measurement development process and in the process of designing the feedback report, the RCT did not include a process to obtain supervisor perceptions of the feedback and its effect on their supervision. This would have helped to inform our interpretations of expected findings, hypotheses about null findings, and explanations for the unexpected finding. Supervisors' experiences of key features of the system (e.g., random sampling of sessions, frequency of feedback, timing of feedback, contents of reports), perceptions of its strengths and weaknesses relative to other sources of feedback they receive routinely, and suggestions for improvement, are

needed to guide further development of useful, valued, and impactful systems. Fourth, as noted above, the process of generating the feedback reports was not fully automated. Fifth, the results of the study do not speak to the value proposition of the SOCS with respect to the implementation and outcomes of MST. Evaluation is needed of the effects of SAF on therapist adherence and youth outcomes, as are precise estimates of the up-front and ongoing costs of maintaining such a system.

Finally, the study took place in the context of a quality assurance and improvement system used to support the implementation of MST. Key features of this system are the collection and review of data on therapist adherence (family-reported), supervisor adherence (therapist-reported), and client outcomes. The system is intended to support a "culture of feedback" (Bickman, 2020), and relative to supervisors in community settings who are newly exposed to EBT supervision, there may have been fewer obstacles related to the overall acceptability of monitoring and feedback or to engaging with a feedback system. To the extent that quality improvement systems, routine outcomes monitoring, and measurement-based care gain purchase in mental health (O'Donohue & Maragakis, 2016), the acceptability and use of feedback may increase, particularly if such feedback is informed by effective A&F interventions. In addition, the SOCS indexed principles and processes specific to MST. However, these principles and processes, as well as the group supervision format, have been used in the supervision of other innovative mental health services for children and families (Atkins et al., 2015; Schoenwald et al., 2013.). Alternatively, SOCS content could be adapted to other EBTs.

Conclusion

The IRT and Rasch modeling approaches to measurement used in this study illuminate distinct attributes of measurement performance and accuracy, accommodate nested data, require relatively low numbers of respondents and items, and may be well suited to advance accurate and feasible measurement in community settings of clinical supervision and other implementation support strategies. Study procedures illuminated the feasibility for community-based clinical supervisors of digitally recording and uploading supervision sessions, and individuals who are neither treatment model experts nor clinicians can be trained to accurately code the supervision sessions. The coded data were used in a web-based audit-and feedback intervention, the SAF. Results of the randomized trial testing the effects of the SAF suggest a single feedback report monthly can affect supervisor performance.

Acknowledgments

Sonja K. Schoenwald is a co-founder and part owner of MST® Services, LLC, which has the exclusive agreement through the Medical University of South Carolina for the transfer of MST technology. She also receives royalties from Guilford Press for published volumes on MST. There is a management plan in place to ensure these conflicts do not jeopardize the objectivity of this research. She did not collect or analyze data for the study. Ashli J. Sheidow is a co-owner of Science to Practice Group, LLC, which provides the training and quality assurance for an adaptation of MST for emerging adults (MST-EA). There is a management plan in place to ensure this conflict does not jeopardize the objectivity of this research. She did not collect or analyze data for the study. Phillippe B. Cunningham is a part owner of Evidence Based Services, Inc., a MST Network Partner Organization. He also receives royalties from Guilford Press for published volumes on MST. There is a management plan in place to ensure these conflicts do not jeopardize the objectivity of this research. He did not collect or analyze data for the study.

Funding for the study was provided by Grant R21/R33MH097000 from the National Institute of Mental Health. The authors wish to thank R21 project coordinator Jennifer Smith Powell and R33 project coordinator Erin McKercher for managing all aspects of the data collection efforts. The authors are grateful to the supervisors in the study, whose dedication to service includes participating in research that might improve it; and, to the leadership of the provider organizations who supported that participation.

Appendix A

Initial Development and Evaluation of the Supervisor Observational Coding System

The Supervisor Observational Coding System (SOCS) was developed in the first study (R21MH097000). The measurement development process included five steps that were based on the *Standards for Educational and Psychological Testing* (*SEPT*; APA, AERA, NCME, & 1999) and associated methods from Item Response Theory (IRT; Wilson, 2005; Stone, 2003; Wolfe & Smith, 2007). The development team for the SOCS included four MST content experts (two MST researchers and two MST Expert trainers), a fidelity measurement consultant, and a measurement development expert. The resulting instrument was pilot tested in the first study, and the instrument was then revised for use in the present study. The psychometric performance of the revised SOCS is detailed extensively in the Results, and the five steps of the initial measurement development process are described next.

Step 1: Define the Purpose of the Instrument and Intended Use of the Scores

The purpose of the instrument was to measure the primary outcome for the experimental supervisor audit-and-feedback (SAF) system that is the focus of this manuscript. Additionally, the instrument and scores were intended for routine monitoring of supervisor fidelity in real-world practice settings. Importantly, the instrument was to be used with audio or video recordings that were rated by trained observational coders. Without separate revision and evaluation efforts, the instrument was not intended for use with self-reports, retrospective reports, or other non-observational reports from other respondents.

Step 2: Define the Main Requirements for the Instrument

The SOCS was intended to be coded in approximately real-time. Most components would be rated for Adherence (i.e., whether the component was delivered), and components that were delivered would also be rated for Competence (i.e., the quality of delivery). Related to this, some components were “always applicable” and therefore would only receive a rating for competence. All components would need to be directly observable from audio-recordings of group supervision sessions. The sessions would be structured with a series of case discussions (typically prioritized by clinical need) among a team of three to four therapists and one supervisor. Accordingly, ratings of individual case discussions were determined to be preferable to providing one set of ratings for the overall supervision session; however, scoring was not necessarily intended to occur at the level of individual case discussions.

Step 3: Define the Components of MST Supervisor Fidelity

This step involved defining a complete list of components, defining rating scale constructs and category labels for adherence and competence, and developing a coding manual for use by the observational coders. Leveraging existing MST supervision materials, candidate components were identified across three theoretical dimensions: Analytical Process (AP), Use of Principles (P), and Structure and Process (SP). A fourth dimension, Delivery Method (DM), was also defined. To ensure that the identified components would be suitable for supervisors with varying levels of adherence and competence, each was located, in a one-to-three-word description, on a hypothetical continuum of supervisors that ranged from novice to expert. This continuum oriented the development team to the concepts of "difficulty" and "ability" which are essential to IRT-based measurement. Each component was located at the point where a supervisor with the given level of adherence or competence would be expected to deliver the component on a consistent basis. Using the information that resulted from this process, a coding manual was developed. Specifically, the coding manual included definitions of Adherence and Competence, a log of decision-rules and modifications, and definitions of each domain and component. For each component, there was a broad definition, definitions specific to Adherence and Competence, a list of terms used by supervisors when delivering the component, examples, counter-examples, and distinctions from similar components. Across the AP, P, SP, and DM domains, the resulting instrument included 40 components, with 13 for AP, 10 for P, 10 for SP, and 7 for DM. For Adherence, each component was rated on a 2-point scale (i.e., 0 = Not Delivered, 1 = Delivered), and components that were delivered were also rated for Competence on a 3-point scale (i.e., 1 = Low, 2 = Moderate, 3 = High). Of note, because the SP components were always applicable, all but one were only rated for Competence.

Step 4: Pilot Test the Coding System

Following procedures approved by the Institutional Review Board of the Medical University of South Carolina, 30 MST supervisors, located in more than 20 sites across the US, recorded weekly supervision sessions for a period of 10 consecutive weeks. A digital recorder was provided by the study, and following each session, the recording was uploaded to a secure server at MUSC via a web-based interface. The trained observational coders, hired and trained for the purpose of this study, were three master's level individuals not involved in MST. The resulting pilot data were analyzed using IRT-based Rasch measurement models, and based on these results, the instrument was revised (see Step 5). There was no evidence of additional dimensionality within the AP, P, SP, or DM domains. The rate of absolute agreement across coders ranged from 78% to 88% for Adherence but was lower for Competence, rating from 39% to 54%. The three-point ordered categorical rating scale for Competence performed as expected with the exception of the DM domain, where only two categories were well-discriminated. For Adherence ratings in the AP, P, and SP domains, the components were well-targeted to the distribution of supervisors, with the components spanning a range of "difficulty" and assessing the full range of supervisor "ability." For Competence ratings, the three-point scale provided good coverage of the supervisor distribution, though supervisors at the highest and lowest levels were not well-targeted. Across domains, four components evidenced unpredictable Adherence ratings, with

five evidencing unpredictable Competence ratings. In each case, the pattern of misfit was suggestive of ambiguous component definitions and thresholds for endorsement.

Step 5. Refine the SOCS for Use in the Second Study

The SOCS was revised based on the psychometric results from the pilot study. The most significant change was that the DM domain was dropped, primarily to reduce coder burden. The revised instrument was comprised of three domains: AP with 10 components, P with 9 components, and SP with 7 components. The final components are reported in Table A1. On the revised instrument, all AP and P components were rated both for Adherence and Competence. For SP, all of the components, with two exceptions, were rated for Competence only.

Table A1

Final SOCS Fidelity Components

		Ratings	
		Adherence	**Competence**
Analytical Process (AP)			
1.	Referral Behaviors	✓	✓
2.	Desired Outcomes of Key Participants	✓	✓
3.	Overarching Goals	✓	✓
4.	Alignment & Engagement	✓	✓
5.	Multisystemic Conceptualization of Fit	✓	✓
6.	Establish Treatment Goals	✓	✓
7.	Prioritize (Drivers, Goals, Interventions)	✓	✓
8.	Specify Interventions to Implement	✓	✓
9.	Measurement Implementation & Effectiveness	✓	✓
10.	Identify Advances & Barriers	✓	✓
Principles (P)			
1.	Finding the Fit	✓	✓
2.	Positive & Strength-Focused	✓	✓
3.	Increasing Responsibility	✓	✓
4.	Present-Focused, Action-Oriented, & Well-Defined	✓	✓
5.	Targeting Sequences	✓	✓
6.	Developmentally Appropriate	✓	✓
7.	Continuous Effort	✓	✓
8.	Evaluation & Accountability	✓	✓
9.	Generalization	✓	✓
Structure and Process (SP)			
1.	Sets Agenda for Session		✓[a]
2.	Obtains Updates on Cases		✓
3.	Identifies Top Clinical Concern	✓	✓
4.	Conveys Urgency for Actions		✓
5.	Facilitates Active Participation of all Therapists	✓	✓

		Ratings	
		Adherence	Competence
6.	Manages Case Discussion		✓
7.	Effectively Manages Time Throughout Session		✓[b]

[a]Only rated for first case discussion.

[b]Only rated for last case discussion.

Appendix B Example Feedback Reports

Figure B1.
Complete Feedback Report

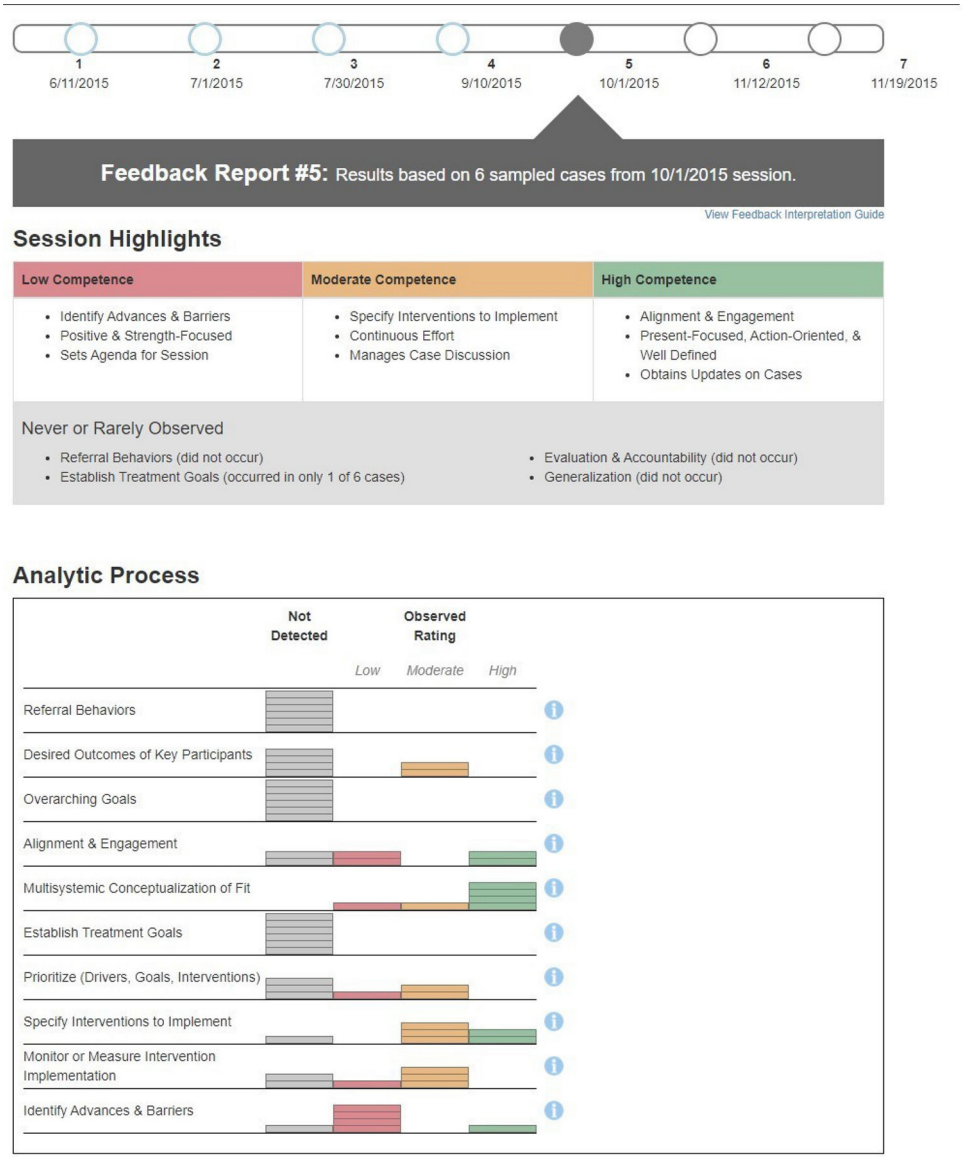

Figure B2.
Analytical Process Section of the Feedback Report

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (1999). Standards for Educational and Psychological Testing.

Atkins MS, Shernoff ES, Frazier SL, Schoenwald SK, Capella E, Marinez-Lora A, Mehta TG, Lakind D, Cua G, Bhaumik R, & Bhaummik D (2015). Redesigning community mental health services for urban children: Supporting schooling to promote mental health. Journal of Consulting and Clinical Psychology, 83, 839–852. 10.1037/a0039661 [PubMed: 26302252]

Bailin A, Bearman SK, & Sale R (2018). Clinical supervision of mental health professionals serving youth: Format and microskills. Administration and Policy in Mental Health and Mental Health Services Research, 45, 800–812. 10.1007/s10488-018-0865-y [PubMed: 29564586]

Bearman SK, Weisz JR, Chorpita BF, Hoagwood K, Ward A, Ugueto AM, Bernstein A, The Research Network on Youth Mental Health (2013). More practice, less preach? The role of supervision processes and therapist characteristics in EBP implementation. Administration and Policy in Mental Health and Mental Health Services Research, 40, 518–529. DOI 10.1007/s10488-013-0485-5 [PubMed: 23525895]

Beidas RS, & Kendall PC (2010). Training therapists in evidence-based practice: A critical review of studies from a systems-contextual perspective. Clinical Psychology: Science and Practice, 17(1), 1–30. Doi:10.1111/j.1468-2850.2009.01187.x [PubMed: 20877441]

Beidas RS, Maclean JC, Fishman J, Dorsey S, Schoenwald SK, Mandell DS, Shea JA, McLeod BD, French MT, Hogue A, Adams DR, Lieberman A, Becker-Haimes M, & Marcus SC (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: Project FACTS study protocol. BMC Psychiatry, 16(323), 1–10. 10.1186/s12888-016-1034-z [PubMed: 26739960]

Bickman L (2000). Our quality-assurance methods aren't so sure. Behavioral Health Tomorrow, 9(3), 41–42.

Bickman L (2020). Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision mental health. Administration and Policy in Mental Health and Mental Health Services Research, 47(5), 795–843. 10.1007/s10488-020-01065-8 [PubMed: 32715427]

Bond TG, & Fox CM (2015). Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge.

Boxmeyer CL, Lochman JE, Powell NR, Windle M, & Wells K (2008). School counselors' implementation of Coping Power in a dissemination field trial: Delineating the range of flexibility within fidelity. Emotional and Behavioral Disorders in Youth, 8, 79–95.

Bradshaw T, Butterworth A, & Mairs H (2007). Does structured clinical supervision during psychosocial intervention education enhance outcome for mental health nurses and the service users they work with? Journal of Psychiatric & Mental Health Nursing, 14, 4–12. 10.1111/j.1365-2850.2007.01021.x [PubMed: 17244000]

Brehaut JC, & Eva KW (2012). Building theories of knowledge translation interventions: Use the entire menu of constructs. Implementation Science, 7(1), 114. 10.1186/1748-5908-7-114 [PubMed: 23173596]

Brunette MF, Cotes RO, de Nesnera A, McHugo G, Dzebisashvili N, Xie H, & Bartels SJ (2018). Use of academic detailing with audit and feedback to improve antipsychotic pharmacotherapy. Psychiatric Services, 69(9), 1021–1028. 10.1176/appi.ps.201700536 [PubMed: 29879874]

Caron EB, & Dozier M (2019). Effects of fidelity-focused consultation on clinicians' implementation: An exploratory multiple baseline design. Administration and Policy in Mental Health and Mental Health Services Research, 46, 445–457. 10.1007/s10488-019-00924-3 [PubMed: 30783903]

Chambers DA, Glasgow R, & Stange K (2013). The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. Implementation Science, 8(1), 117. doi:10.1186/1748-5908-8-117 [PubMed: 24088228]

Charlton C, Rasbash J, Browne W, Healy M, & Cameron B (2019). MLwiN (Version 3.03) [Computer software and manual]. Centre for Multilevel Modelling. http://www.bristol.ac.uk/cmm/software/mlwin/

Chorpita BF, & Regan J (2009). Dissemination of effective mental health treatment procedures: Maximizing the return on a significant investment. Behaviour Research and Therapy, 47, 990–993. 10.1016/j.brat.2009.07.002 [PubMed: 19632669]

Colquhoun H, Michie S, Sales S, Ivers N, Grimshaw JM, Carroll K, Chalifoux M, Eva K & Brehaut J (2017). Reporting and design elements of audit and feedback interventions: a secondary review. BMJ Quality and Safety 26, 54–60. Doi:10.1136/bmjqs-2015-005004 [PubMed: 26811541]

Colquhoun HL, Carroll K, Eva KW, Grimshaw JG, Ivers N, Michie S, & Brehaut JC (2021). Informing the research agenda for optimizing audit and feedback interventions: Results of a prioritization exercise. BMC Medical Research Methodology, 21(1), 20. 10.1186/s12874-020-01195-5 [PubMed: 33435873]

Collyer H, Eisler I, & Woolgar M (2020). Systematic literature review and meta-analysis of the relationship between adherence, competence and outcome in psychotherapy for children and adolescents. European Child & Adolescent Psychiatry, 29(4), 417–431. 10.1007/s00787-018-1265-2 [PubMed: 30604132]

Dorsey S, Kerns SEU, Lucid L, Pullmann MD, Harrison JP, Berliner L, Thompson K, & Deblinger E (2018). Objective coding of content and techniques in workplace-based supervision of an EBT in public mental health. Implementation Science, 13(1), 19. 10.1186/s13012-017-0708-3 [PubMed: 29368656]

Dorsey S, Pullmann MD, Deblinger E, Berliner L, Kerns SE, Thompson K, Unützer J, Weisz JR, & Garland AF (2013). Improving practice in community-based settings: A randomized trial of supervision – study protocol. Implementation Science, 8, 89. 10.1186/1748-5908-8-89 [PubMed: 23937766]

Dorsey S, Pullmann MD, Kerns SEU, Jungbluth N, Meza R, Thompson K, & Berliner L (2017). The juggling act of supervision in community mental health: Implications for supporting evidence-based treatment. Administration and Policy in Mental Health and Mental Health Services Research, 44, 838–852. 10.1007/s10488-017-0796-z [PubMed: 28315076]

Farmer CC, Mitchell KS, Parker-Guilbert K, & Galovski TE (2016). Fidelity to the cognitive processing therapy protocol: Evaluation of critical elements. Behavior Therapist, 48(2), 195–206. 10.1016/j.beth.2016.02.009

Fixsen DL, Naoom SF, Blasé KA, Friedman RM, & Wallace F (2005). Implementation research: A synthesis of the literature. University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Glasgow RE, & Riley WT (2013). Pragmatic measures: What they are and why we need them. American Journal of Preventive Medicine, 45(2), 237–243. 10.1016/j.amepre.2013.03.010 [PubMed: 23867032]

Gleacher AA, Olin SS, Nadeem E, Pollock M, Ringle V, Bickman L, Douglas S, & Hoagwood K (2016). Implementing a measurement feedback system in community mental health clinics: A case study of multilevel barriers and facilitators. Administration and Policy in Mental Health and Mental Health Services Research, 43(3), 426–440. 10.1007/s10488-015-0642-0. [PubMed: 25735619]

Henggeler SW, Pickrel SG, & Brondino M J. (1999). Multisystemic treatment of substance abusing and dependent delinquents: Outcomes, treatment fidelity, and transportability. Mental Health Services Research, 1, 171–184. [PubMed: 11258740]

Henggeler SW, Schoenwald SK, Borduin CM, Rowland MD, & Cunningham PB (2009). Multisystemic Therapy for antisocial behavior in children and adolescents (2nd ed.). The Guilford Press.

Henggeler SW, Melton GB, Brondino MJ, Scherer DG, & Hanley JH (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. Journal of Consulting and Clinical Psychology, 65, 821–833. [PubMed: 9337501]

Hogue A, Liddle HA, & Rowe C (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. Psychotherapy: Theory, Research, Practice, Training, 33, 332–345. Doi:10.1037/0033-3204.33.2.332

Hogue A, Henderson CE, Dauber S, Barajas PC, Fried A, & Liddle HA (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. Journal of Consulting and Clinical Psychology, 76(4), 544–555. 10.1037/0022-006X.76.4.544 [PubMed: 18665684]

Huey SJ, Henggeler SW, Brondino MJ, & Pickrel SG (2000). Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. Journal of Consulting and Clinical Psychology, 68, 451–467. [PubMed: 10883562]

Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, O'Brien MA, Johansen M, Grimshaw J, & Oxman AD (2012). Audit and feedback: Effects on professional practice and healthcare outcomes. Cochrane Database Systematic Review, 6, CD000259. 10.1002/14651858.CD000259.pub3

Ivers N, Sales A, Colquhoun H, Michie S, Foy R, Francis JJ, & Grimshaw JM (2014). No more 'business as usual' with audit and feedback interventions: Towards an agenda for a reinvigorated intervention. Implementation Science, 9, 14. 10.1186/1748-5908-9-14 [PubMed: 24438584]

Kamata A (2001). Item analysis by the hierarchical generalized linear model. Journal of Educational Measurement, 38, 79–93. https://doi.org/10.1111%2Fj.1745-3984.2001.tb01117.x

Lambert MJ, & Harmon KL (2018). The merits of implementing routine outcome monitoring in clinical practice. Clinical Psychology: Science and Practice, 25(4), e12268. 10.1111/cpsp.12268

Lambert MJ, Whipple JL, & Kleinstäuber M (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. Psychotherapy, 55(4), 520–537. 10.1037/pst0000167 [PubMed: 30335463]

Landis-Lewis Z, Kononowech J, Scott WJ, Hogikyan RV, Carpenter JG, Periyakoil VS, Miller SC, Levy C, Ersek M, & Sales A (2020). Designing clinical practice feedback reports: Three steps illustrated in Veterans Health Affairs long-term care facilities and programs. Implementation Science, 15, 7. 10.1186/s13012-019-0950-y [PubMed: 31964414]

Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, Kassab H, Hoffman M, Scott K, Lyon A, Douglas S, Simon G, & Kroenke K (2019). Implementing measurement-based care in behavioral health: A review. JAMA Psychiatry, 76(3), 324–335. 10.1001/jamapsychiatry.2018.3329 [PubMed: 30566197]

Linacre JM (1994). Many-facet Rasch measurement. MESA Press.

Linacre JM (2002). Optimizing rating scale category effectiveness. Journal of Applied Measurement, 3, 85–106. [PubMed: 11997586]

Linacre JM (2019a). FACETS Rasch measurement computer program [Computer software and manual]. https://www.winsteps.com/facets.htm

Linacre JM (2019b). WINSTEPS Rasch measurement computer program [Computer software and manual]. https://www.winsteps.com/facets.htm

Linacre JM, & Wright BD (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370.

Lochman JE, Boxmeyer C, Powell N, Qu L, Wells K, & Windle M (2009). Dissemination of the Coping Power program: Importance of intensity of counselor training. Journal of Consulting and Clinical Psychology, 77, 397–409. 10.1037/a0014514 [PubMed: 19485582]

Lyon AR, Lewis CC, Boyd MR, Hendrix E, & Liu F (2016). Capabilities and characteristics of digital measurement feedback systems: Results from a comprehensive review. Administration and Policy in Mental Health and Mental Health Services Research, 43(3), 441–466. 10.1007/s10488-016-0719 [PubMed: 26860952]

Maas CJM, & Hox JJ (2005). Sufficient sample sizes for multilevel modeling. Methodology, 1, 86–92. 10.1027/1614-2241.1.3.86

Myford CM, & Wolfe EW (2003). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I. Journal of Applied Measurement, 4, 386–422. [PubMed: 14523257]

Newman CF, Reiser RP, & Milne DL (2016). Supporting our supervisors: a summary and discussion of the special issue on CBT supervision. The Cognitive Behavior Therapist, 9, e29doi:10.1017/S1754470X16000106

Ogden T, & Hagen KA (2006). Multisystemic therapy of serious behavior problems in youth: Sustainability of therapy effectiveness two years after intake. Child and Adolescent Mental Health, 11(3), 142–149. [PubMed: 32811000]

O'Donohue W, & Maragakis A, Eds. (2016). Quality Improvement in Behavioral Health. Springer.

Perepletchikova F, Treat TA, & Kazdin AE (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. Journal of Consulting and Clinical Psychology, 75, 829–841. 10.1037/0022-006X.75.6.829 [PubMed: 18085901]

Rasch G (1980). Probabilistic models for some intelligence and achievement tests. University of Chicago Press.

Raudenbush SW, & Bryk AS (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Sage Publications.

Raudenbush SW, Bryk AS, & Congdon R (2013). HLM 7: Hierarchical linear & nonlinear modeling (version 7.00) [Computer software & manual]. Scientific Software International. https://ssicentral.com/index.php/products/hlm-general/

Ravand H (2015). Item Response Theory using hierarchical generalized linear models. Practical Assessment, Research, & Evaluation, 20(7), 1–17. 10.7275/s4n1-kn37

Roth AD, Pilling S, & Turner J (2010). Therapist training and supervision in clinical trials: Implications for clinical practice. Behavioral and Cognitive Psychotherapy, 38(3), 291–302. Doi:10.1017/S13524658100000068

Saldana L, Chamberlain P, & Chapman J (2016). A supervisor-targeted implementation approach to promote system change: The R^3 Model. Administration and Policy in Mental Health and Mental Health Services Research, 43(6), 879–892. 10.1007/s10488-016-0730-9 [PubMed: 27003137]

Sale R, Bearman SK, Woo R, & Baker N (2021). Introducing a measurement feedback system for youth mental health: Predictors and impact of implementation in a community agency. Administration and Policy in Mental Health and Mental Health Services Research, 48(2), 327–342. 10.1007/s10488-020-01076-5 [PubMed: 32809082]

Schoenwald SK (1998). MST Personnel Data Inventory. Family Services Research Center, Medical University of South Carolina.

Schoenwald SK (2016). The Multisystemic Therapy® quality assurance/quality improvement system. In O'Donohue W & Maragakis A,(Eds.), Quality Improvement and Behavioral Health (pp. 169–192). Switzerland: Springer International Publishing Switzerland

Schoenwald SK, Chapman JE, Kelleher K, Hoagwood KE, Landsverk J, Stevens J, Glisson C, Rolls-Reutz J, and The Research Network on Youth Mental Health (2008). A survey of the infrastructure for children's mental health services: Implications for the implementation of empirically supported treatments (ESTs). Administration and Policy in Mental Health and Mental Health Services Research, 35, 84–97. 10.1007/s10488-007-0147 [PubMed: 18000750]

Schoenwald SK, & Garland AF (2013). A review of treatment adherence measurement methods. Psychological Assessment, 25, 146–156. 10.1037/a0029715 [PubMed: 22888981]

Schoenwald SK, Garland AF, Chapman JE, Frazier SL, Sheidow AJ, & Southam-Gerow MA (2011). Toward the effective and efficient measurement of implementation fidelity. Administration and Policy in Mental Health and Mental Health Services Research, 38, 32–43. 10.1007/s10488-010-0321-0 [PubMed: 20957425]

Schoenwald SK, Mehta TG, Frazier SL, & Shernoff ES (2013). Clinical supervision in effectiveness and implementation research. Clinical Psychology: Science and Practice, 20, 44–59. 10.1080/14733140601185274

Schoenwald SK, Sheidow AJ, & Chapman JE (2009). Clinical supervision in treatment transport: Effects on adherence and outcomes. Journal of Consulting and Clinical Psychology, 77(3), 410–421. 10.1037/a0013788 [PubMed: 19485583]

Singer JD, & Willett JB (2003). Applied longitudinal data analysis: Modeling change and event occurrence. Oxford University Press.

Smith EV Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement, 3, 205–231. [PubMed: 12011501]

Southam-Gerow M, Chapman JE, Martinez RG, McLeod BD, Hogue A, Weisz JR, & Kendall PC (2021). Are therapist adherence and competence related to clinical outcomes in cognitive-behavioral treatment for youth anxiety? Journal of Consulting and Clinical Psychology, 89 (3), 188–199. 10.1037/ccp0000538 [PubMed: 33829807]

Stirman SW, Calloway A, Toder K, Miller CJ, DeVito AK, Meisel SN, Xhezo R, Evans AC, Beck AT, & Crits-Christoph P (2013). Modifications to cognitive therapy by community mental health providers: Implications for effectiveness and sustainability. Psychiatric Services, 64(10), 1056–1059. 10.1176/appi.ps.201200456 [PubMed: 24081406]

Stirman SW, Finley EP, Shields N, Cook J, Haine-Schlagel R, Burgess JF, Dimeff L, Koerner K, Suvak M, Gutner CA, Gagnon D, Masina T, Beristianos M, Mallard K, Ramirez V, & Monson C (2017). Improving and sustaining delivery of CPT for PTSD in mental health systems: A cluster

randomized trial. Implementation Science, 12(1), 32. 10.1186/s13012-017-0544-5 [PubMed: 28264720]

Stirman SW, Kimberly JR, Calloway A, Cook N, Castro F, & Charns MP (2012). The sustainability of new programs and interventions: A review of the empirical literature and recommendations for future research. Implementation Science, 7(1), 17. 10.1186/1748-5908-7-17 [PubMed: 22417162]

Stirman SW, Marques L, Creed TA, Cassidy AG, DeRubeis R, Barnett PG, Kuhn E, Suvak M, Owen J, Vogt D, Jo B, Schoenwald S, Johnson C, Mallard K, Beristianos M, & La Bash H (2018). Leveraging routine clinical materials and mobile technology to assess CBT fidelity: The Innovative Methods to Assess Psychotherapy Practices (imAPPP) study. Implementation Science, 13(1), 69. 10.1186/s13012-018-0756-3 [PubMed: 29789017]

Stirman SW, Shields N, Deloriea J, Landy MSH, Belus JM, Maslej MM, & Monson CM (2013). A randomized controlled dismantling trial of post-workshop consultation strategies to increase effectiveness and fidelity to an evidence-based psychotherapy for posttraumatic stress disorder. Implementation Science, 8, 82. 10.1186/1748-5908-8-82 [PubMed: 23902798]

Weisz JR, Ng MY, & Bearman SK (2014). Odd couple? Reenvisioning the relation between science and practice in the dissemination-implementation era. Clinical Psychological Science, 2(1), 58–74. 10.1177/2167702613501307

Wright BD, & Masters G (1982). Rating scale analysis. MESA Press.

Wright BD, & Mok M (2000). Rasch models overview. Journal of Applied Measurement, 1, 83–106 [PubMed: 12023559]

Yeaton WH, & Sechrest L (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. Journal of Consulting and Clinical Psychology, 49, 156–167. [PubMed: 7217482]

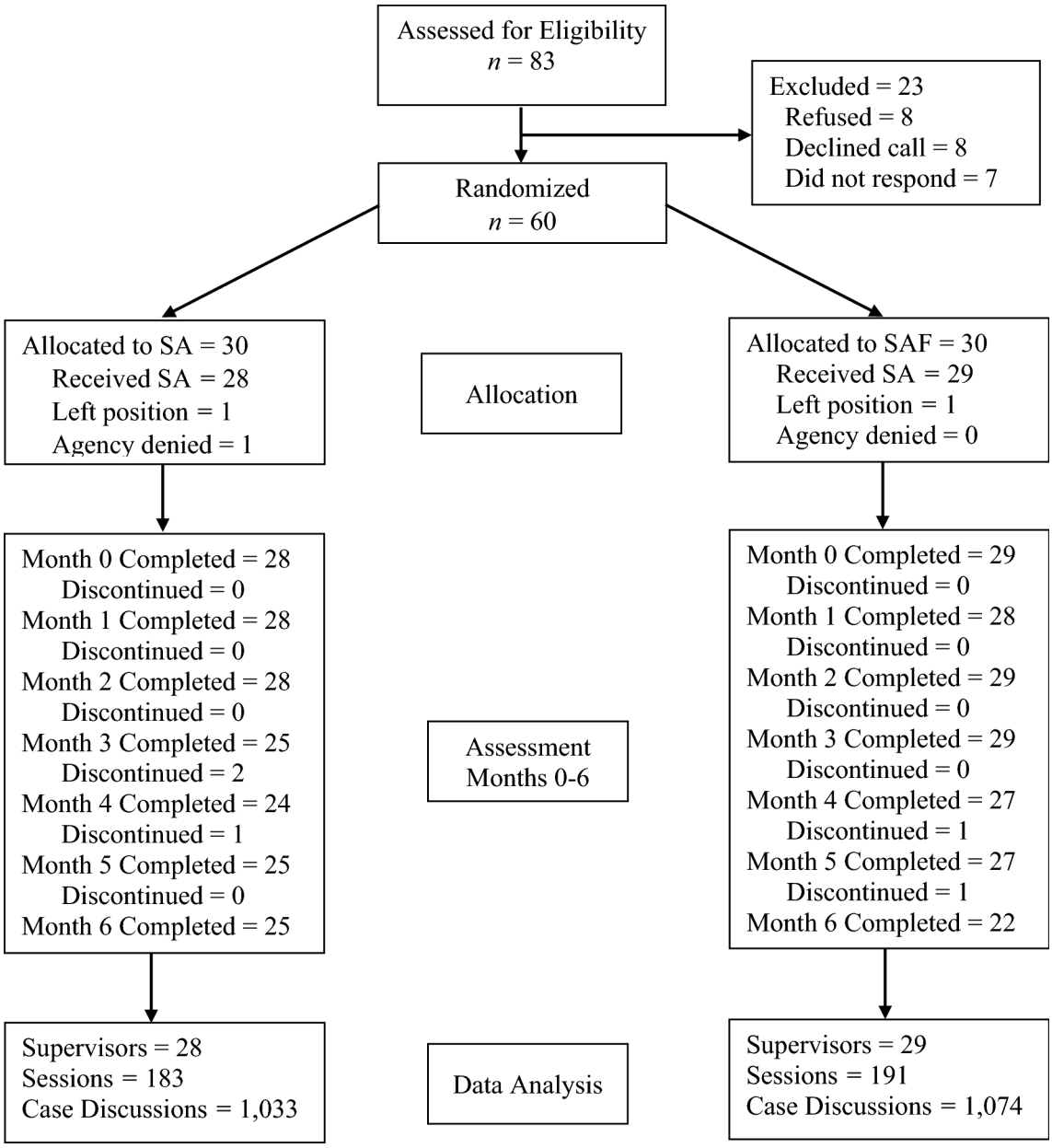

Figure 1.
CONSORT flow diagram

Table 1

Variance components, variance proportions, and reliability estimates from unconditional three-level mixed-effects regression models for supervisor adherence and competence outcomes in each domain

	Adherence						Competence					
	Variance		Prop.[a]		Reliability[b]		Variance		Prop.[a]		Reliability[b]	
	Rasch[c]	Raw[d]	Rasch[c]	Raw[d]	Rasch[c]	Raw[d]	Rasch[c]	Raw[d]	Rasch[c]	Raw[d]	Rasch[c]	Raw[d]
Analytical Process												
Case	0.329	0.023	0.46	0.75			0.100	0.151	0.18	0.67		
Session	0.058	0.002	0.08	0.05	0.50	0.27	0.312	0.061	0.57	0.27	0.94	0.69
Supervisor	0.323	0.006	0.45	0.20	0.95	0.87	0.137	0.013	0.25	0.06	0.73	0.48
MST Principles												
Case	0.182	0.018	0.38	0.76			0.068	0.147	0.53	0.69		
Session	0.104	0.003	0.22	0.11	0.76	0.44	0.259	0.048	0.14	0.23	0.95	0.64
Supervisor	0.193	0.003	0.40	0.13	0.90	0.78	0.159	0.017	0.33	0.08	0.79	0.59
Structure & Process												
Case							*e*	0.069	*e*	0.52		
Session							0.094	0.022	0.23	0.16	*e*	0.64
Supervisor							0.320	0.042	0.77	0.31	0.96	0.89

[a] Prop. is the proportion of total outcome variance attributable to the respective level of nesting (i.e., case, session, supervisor).

[b] The multilevel reliability estimate reflects both precision and reliability, reflecting the reliability of scores at the respective level of measurement.

[c] Logit-based Rasch measures (i.e., “scores”) were computed using empirical Bayes residuals from Rasch-equivalent hierarchical generalized linear measurement models.

[d] Average scores.

[e] For the SP Competence outcome based on Rasch measures, the variance in case discussions (i.e., within-sessions) was near 0. Accordingly, the individual case discussion scores were removed, and a two-level formulation was used with repeated Rasch measures for sessions (level-1) nested within supervisors (level-2).

Table 2

Models for linear change over time for supervisor adherence and competence outcomes in each domain

	Fixed Effect Estimates					
	Rasch Measure [a]			Raw Score [b]		
	Est.	*SE*	*p*	Est.	*SE*	*p*
AP Adherence						
Intercept	0.102	0.116	.382	0.377	0.018	<.001
SAF	−0.152	0.163	.354	−0.035	0.025	.166
Linear	−0.033	0.013	.010	−0.007	0.003	.009
SAF × Linear	0.047	0.018	.009	0.010	0.004	.007
Contrast [c]						
Linear (SAF)	0.014	0.013	.259	0.003	0.003	.231
AP Competence						
Intercept	0.044	0.093	.637	1.971	0.040	<.001
SAF	−0.167	0.130	.204	−0.072	0.057	.207
Linear	−0.019	0.024	.420	−0.008	0.013	.544
SAF × Linear	0.065	0.034	.058	0.029	0.018	.105
Contrast [c]						
Linear (SAF)	0.046	0.024	.051	0.022	0.012	.080
P Adherence						
Intercept	0.028	0.096	.771	0.353	0.010	<.001
SAF	0.005	0.135	.973	−0.002	0.018	.901
Linear	−0.018	0.014	.208	−0.003	0.003	.213
SAF × Linear	0.014	0.020	.481	0.003	0.004	.489
Contrast [c]						
Linear (SAF)	−0.004	0.014	>.500	−0.001	0.003	>.500
P Competence						
Intercept	0.013	0.101	.896	1.983	0.043	<.001
SAF	−0.010	0.142	.943	−0.007	0.060	.913
Linear	−0.015	0.019	.450	−0.006	0.010	.546
SAF × Linear	0.023	0.027	.404	0.012	0.014	.392
Contrast [c]						
Linear (SAF)	0.008	0.019	>.500	0.006	0.010	>.500
SP Competence [d]						
Intercept	0.024	0.120	.844	2.079	0.048	<.001
SAF	−0.080	0.168	.634	−0.057	0.067	.401
Linear	−0.013	0.015	.376	−0.011	0.009	.211
SAF × Linear	0.037	0.021	.076	0.029	0.012	.019
Contrast [c]						

Linear (SAF)	0.024	0.015	.091	0.018	0.009	.030

	Variance Components					
	Rasch Measure [a]			Raw Score [b]		
	Var	*SD*	*p*	Var	*SD*	*p*
AP Adherence						
Error	0.329	0.574		0.023	0.151	
Session	0.055	0.235	<.001	0.001	0.037	<.001
Linear						
Supervisor	0.324	0.569	<.001	0.006	0.078	<.001
AP Competence						
Error	0.100	0.316		0.151	0.389	
Session	0.292	0.540	<.001	0.055	0.235	<.001
Linear	0.003	0.058	.055	0.001	0.033	.028
Supervisor	0.092	0.304	.001	0.006	0.078	.113
P Adherence						
Error	0.182	0.427		0.018	0.135	
Session	0.103	0.321	<.001	0.003	0.050	<.001
Linear						
Supervisor	0.193	0.439	<.001	0.003	0.056	<.001
P Competence						
Error	0.068	0.261		0.147	0.383	
Session	0.258	0.508	<.001	0.048	0.219	<.001
Linear						
Supervisor	0.158	0.397	<.001	0.017	0.129	<.001
SP Competence						
Error	0.082	0.286		0.069	0.262	
Session	*d*	*d*		0.017	0.129	<.001
Linear	0.003	0.052	<.001	0.001	0.030	<.001
Supervisor	0.361	0.601	<.001	0.051	0.226	<.001

Note. The *T*-ratio test statistic (not reported) was computed as Est./*SE*. Est. = Estimate; *SE* = Standard Error.

[a] Logit-based Rasch measures (i.e., "scores") were computed using empirical Bayes residuals from Rasch-equivalent hierarchical generalized linear measurement models.

[b] Average scores.

[c] Planned contrast for linear slope significance in the SAF condition.

[d] For the SP Competence outcome based on Rasch measures, the variance in case discussions (i.e., within-sessions) was near 0. Accordingly, the individual case discussion scores were removed, and a two-level formulation was used with repeated Rasch measures for sessions (level-1) nested within supervisors (level-2). As such, the error variance for this model reflects session-level outcome variance.

Table 3

Models for change between baseline and each later month for supervisor adherence and competence outcomes in each domain

	Fixed Effects					
	Rasch Measure [a]			Raw Score [d]		
	Est.	*SE*	*p*	Est.	*SE*	*p*
AP Adherence						
Intercept	0.143	0.124	.254	0.386	0.020	<.001
SAF	−0.226	0.175	.200	−0.051	0.028	.078
Month 1	−0.089	0.089	.323	−0.020	0.019	.306
SAF × Month 1	0.132	0.126	.297	0.030	0.027	.276
Month 2	−0.166	0.090	.067	−0.037	0.020	.060
SAF × Month 2	0.263	0.126	.038	0.057	0.027	.039
Month 3	−0.133	0.092	.150	−0.027	0.020	.177
SAF × Month 3	0.273	0.128	.033	0.057	0.028	.041
Month 4	−0.131	0.093	.083	−0.037	0.020	.063
SAF × Month 4	0.195	0.129	.132	0.047	0.028	.093
Month 5	−0.194	0.094	.040	−0.038	0.200	.062
SAF × Month 5	0.294	0.130	.025	0.062	0.028	.028
Month 6	−0.231	0.093	.013	−0.054	0.020	.008
SAF × Month 6	0.343	0.133	.011	0.077	0.029	.009
Contrasts [c]						
Month 1 (SAF)	0.043	0.089	.236	0.010	0.019	>.500
Month 2 (SAF)	0.097	0.088	.268	0.020	0.019	.303
Month 3 (SAF)	0.140	0.089	.109	0.030	0.019	.114
Month 4 (SAF)	0.033	0.090	>.500	0.010	0.019	>.500
Month 5 (SAF)	0.100	0.090	.269	0.024	0.020	.216
Month 6 (SAF)	0.112	0.096	.241	0.023	0.021	.269
AP Competence						
Intercept	0.035	0.127	.784	1.947	0.059	<.001
SAF	0.037	0.178	.835	0.042	0.082	.610
Month 1	0.043	0.150	.773	0.066	0.078	.399
SAF × Month 1	−0.246	0.211	.245	−0.177	0.110	.109
Month 2	−0.005	0.151	.972	0.020	0.079	.797
SAF × Month 2	−0.287	0.211	.175	−0.129	0.110	.242
Month 3	−0.162	0.155	.297	−0.046	0.080	.566
SAF × Month 3	0.069	0.214	.746	−0.002	0.111	.989
Month 4	−0.073	0.157	.643	−0.022	0.081	.787
SAF × Month 4	−0.017	0.217	.936	−0.030	0.112	.792
Month 5	−0.081	0.159	.612	0.011	0.081	.896
SAF × Month 5	0.179	0.219	.414	0.035	0.113	.757

Month 6	−0.062	0.155	.690	−0.009	0.080	.896
SAF × Month 6	0.267	0.223	.233	0.115	0.116	.320
Contrasts [c]						
Month 1 (SAF)	−0.203	0.149	.170	−0.111	0.078	.149
Month 2 (SAF)	−0.292	0.147	.044	−0.108	0.076	.152
Month 3 (SAF)	−0.093	0.147	>.500	−0.048	0.077	>.500
Month 4 (SAF)	−0.090	0.150	>.500	−0.051	0.078	>.500
Month 5 (SAF)	0.098	0.151	>.500	0.045	0.079	>.500
Month 6 (SAF)	0.205	0.160	.198	0.106	0.083	.197
P Adherence						
Intercept	0.072	0.107	.503	0.362	0.017	<.001
SAF	−0.074	0.150	.623	−0.020	0.024	.420
Month 1	−0.085	0.096	.375	−0.019	0.020	.336
SAF × Month 1	0.058	0.136	.668	0.017	0.028	.537
Month 2	−0.134	0.097	.169	−0.029	0.020	.145
SAF × Month 2	0.288	0.135	.034	0.059	0.028	.035
Month 3	−0.050	0.101	.622	−0.009	0.020	.662
SAF × Month 3	0.116	0.138	.403	0.024	0.028	.388
Month 4	−0.226	0.100	.025	−0.046	0.020	.025
SAF × Month 4	0.191	0.139	.169	0.041	0.028	.153
Month 5	0.026	0.102	.801	0.007	0.021	.751
SAF × Month 5	−0.060	0.140	.668	−0.010	0.029	.730
Month 6	−0.196	0.010	.051	−0.041	0.020	.043
SAF × Month 6	0.232	0.143	.106	0.049	0.029	.097
Contrasts [c]						
Month 1 (SAF)	−0.027	0.096	>.500	−0.002	0.020	>.500
Month 2 (SAF)	0.154	0.094	.099	0.030	0.019	.119
Month 3 (SAF)	0.066	0.095	>.500	0.015	0.020	>.500
Month 4 (SAF)	−0.034	0.096	>.500	−0.005	0.020	>.500
Month 5 (SAF)	−0.035	0.097	>.500	−0.003	0.020	>.500
Month 6 (SAF)	0.036	0.103	.125	0.007	0.021	>.500
P Competence						
Intercept	−0.004	0.122	.977	1.943	0.055	<.001
SAF	0.081	0.171	.367	0.070	0.078	.370
Month 1	−0.042	0.136	.757	0.058	0.071	.419
SAF × Month 1	0.104	0.191	.585	−0.022	0.100	.829
Month 2	0.173	0.136	.204	0.100	0.072	.162
SAF × Month 2	−0.462	0.190	.016	−0.224	0.100	.025
Month 3	−0.190	0.140	.177	−0.056	0.073	.444
SAF × Month 3	0.111	0.193	.565	0.002	0.101	.981
Month 4	−0.078	0.142	.581	0.030	0.074	.679
SAF × Month 4	0.021	0.196	.915	−0.058	0.102	.569
Month 5	0.022	0.141	.875	0.049	0.074	.507

SAF × Month 5	−0.078	0.196	.689	−0.063	0.103	.542
Month 6	−0.099	0.140	.479	−0.032	0.074	.663
SAF × Month 6	0.209	0.202	.301	0.107	0.106	.312
Contrasts [c]						
Month 1 (SAF)	0.062	0.135	>.500	0.036	0.071	>.500
Month 2 (SAF)	−0.289	0.133	.028	−0.124	0.070	.072
Month 3 (SAF)	−0.078	0.133	>.500	−0.054	0.070	>.500
Month 4 (SAF)	−0.057	0.136	>.500	−0.028	0.071	>.500
Month 5 (SAF)	−0.056	0.136	>.500	−0.014	0.071	>.500
Month 6 (SAF)	0.109	0.145	>.500	0.075	0.076	>.500
SP Competence [d]						
Intercept	0.104	0.122	.396	2.127	0.051	>.001
SAF	−0.095	0.171	.579	−0.082	0.071	.252
Month 1	−0.148	0.080	.067	−0.091	0.047	.054
SAF × Month 1	0.080	0.113	.482	0.074	0.067	.266
Month 2	−0.104	0.080	.197	−0.073	0.047	.125
SAF × Month 2	−0.049	0.113	.667	0.023	0.066	.734
Month 3	−0.237	0.083	.005	−0.154	0.049	.002
SAF × Month 3	0.323	0.115	.005	0.241	0.067	<.001
Month 4	−0.126	0.085	.137	−0.086	0.049	.080
SAF × Month 4	0.156	0.117	.185	0.131	0.068	.054
Month 5	−0.055	0.084	.514	−0.049	0.049	.321
SAF × Month 5	0.143	0.116	.221	0.137	0.068	.045
Month 6	−0.159	0.084	.058	−0.108	0.049	.028
SAF × Month 6	0.227	0.120	.059	0.178	0.070	.012
Contrasts [c]						
Month 1 (SAF)	−0.068	0.080	>.500	−0.017	0.047	>.500
Month 2 (SAF)	−0.153	0.079	.050	−0.050	0.046	.276
Month 3 (SAF)	0.086	0.079	.276	0.087	0.047	.059
Month 4 (SAF)	0.030	0.081	>.500	0.045	0.047	>.500
Month 5 (SAF)	0.088	0.081	.276	0.088	0.047	.059
Month 6 (SAF)	0.069	0.086	>.500	0.070	0.051	.162

	Variance Components					
	Unconditional			Growth		
	Var	*SD*	*p*	Var	*SD*	*p*
AP Adherence						
Error	0.329	0.574		0.023	0.151	
Session	0.054	0.232	<.001	0.001	0.036	<.001
Supervisor	0.324	0.569	<.001	0.006	0.078	<.001
AP Competence						
Error	0.100	0.316		0.151	0.389	

Session	0.300	0.545	<.001	0.058	0.240	<.001
Supervisor	0.139	0.372	<.001	0.013	0.114	<.001
P Adherence						
Error	0.182	0.427		0.018	0.135	
Session	0.097	0.311	<.001	0.002	0.047	<.001
Supervisor	0.193	0.439	<.001	0.003	0.056	<.001
P Competence						
Error	0.068	0.261		0.147	0.384	
Session	0.245	0.495	<.001	0.044	0.211	<.001
Supervisor	0.158	0.397	<.001	0.017	0.130	<.001
SP Competence						
Error	0.091	0.301		0.069	0.262	
Session	*d*	*d*		0.019	0.138	<.001
Supervisor	0.326	0.571	<.001	0.042	0.204	<.001

Note. The *T*-ratio test statistic (not reported) was computed as Est./*SE*. Est. = Estimate; *SE* = Standard Error.

[a] Logit-based Rasch "scores" computed using empirical Bayes residuals from Rasch-equivalent hierarchical generalized linear measurement models.

[b] Average scores.

[c] Planned contrast for linear slope significance in the SAF condition.

[d] For the SP Competence outcome based on Rasch measures, the variance in case discussions (i.e., within-sessions) was near 0. Accordingly, the individual case discussion scores were removed, and a two-level formulation was used with repeated Rasch measures for sessions (level-1) nested within supervisors (level-2). As such, the error variance for this model reflects session-level outcome variance.