Short Report

Implementation Research & Practice

Implementation Research and Practice
Volume 3: Jan-Dec 2022 1–10

DOI: 10.1177/26334895221135263
journals.sagepub.com/home/irp
SAGE

Improving the feasibility of fidelity measurement for community-based quality assurance: Partial- versus full-session observations of supervisor adherence and competence

Jason E. Chapman[1], Zoe M. Alley[1] and Sonja K. Schoenwald[1]

Abstract

Background

Clinical supervision is a common quality assurance method for supporting the implementation and sustainment of evidence-based interventions (EBIs) in community mental health settings. However, assessing and supporting supervisor fidelity requires efficient and effective measurement methods. This study evaluated two observational coding approaches that are potentially more efficient than coding full sessions: a randomly selected 15-min segment and the first case discussion of the session.

Method

Data were leveraged from a randomized trial of an Audit and Feedback (A&F) intervention for supervisor Adherence and Competence. Supervisors ($N = 57$) recorded and uploaded weekly group supervision sessions for 7 months, with one session observationally coded each month ($N = 374$). Of the coded sessions, one was randomly selected for each supervisor, and a random 15-min segment was coded. Additionally, the first case discussion was coded for the full sample of sessions.

Results

Across all models (and controlling for the proportion of the session covered by the partial observation), Adherence and Competence scores from partial observations were positively and significantly associated with scores from full sessions. In all cases, partial observations were most accurate when the level of Adherence and Competence was moderate. At lower levels, partial observations were underestimates, and at higher levels, they were overestimates.

Conclusions

The results suggest that efficient observational measurement can be achieved while retaining a general level of measurement effectiveness. Practically, first-case discussions are easier to implement, whereas 15-min segments have fewer potential threats to validity. Evaluation of resource requirements is needed, along with determining whether A&F effects are retained if feedback is based on partial observations. Nevertheless, more efficient observational coding could increase the feasibility of routine fidelity monitoring and quality assurance strategies, including A&F, which ultimately could support the implementation and sustainment of effective supervision practices and EBIs in community practice settings.

[1]Oregon Social Learning Center, Eugene, OR, USA

Corresponding author:
Jason E. Chapman, Oregon Social Learning Center, 10 Shelton McMurphey Blvd, Eugene, OR 97401, USA.
Email: jasonc@oslc.org

Plain Language Summary: When delivering evidence-based mental health interventions in community-based practice settings, a common quality assurance method is clinical supervision. To support supervisors, assessment methods are needed, and those methods need to be both efficient and effective. Ideally, supervision sessions would be recorded, and trained coders would rate the supervisor's use of specific strategies. In most settings, though, this requires too many resources. The present study evaluated a more efficient approach. The data came from an existing randomized trial of an Audit and Feedback intervention for enhancing supervisor Adherence and Competence. This included 57 supervisors and 374 sessions across seven months of monitoring. Instead of rating full supervision sessions, a more efficient approach was to have coders rate partial sessions. Two types of partial observations were considered: a randomly selected 15-minute segment of the session and the first case discussion of the session. The aim was to see if partial observations and full observations led to similar conclusions about Adherence and Competence. In all cases, they did. The scores were most similar for sessions with moderate levels of Adherence and Competence. If Adherence and Competence were low, partial observations were underestimates, but if they were high, partial observations were overestimates. Observing partial sessions is more efficient, but in terms of accuracy, the benefits and limitations should be evaluated in light of how the scores will be used. Additionally, future research should consider whether Audit and Feedback interventions have the same effect if feedback is based on observations of partial sessions.

Keywords
clinical supervision, fidelity measurement, adherence and competence, audit and feedback interventions, quality assurance

Introduction

Audit and Feedback (A&F) interventions have been effective in aligning the practices of health care professionals with the guidelines of evidence-based care (Colquhoun et al., 2021). A&F interventions summarize data about specific aspects of practice over a specified period of time and feed that information back to practitioners. However, scant research has focused on the use and effects of A&F in mental health care, and an area ripe for evaluation is the measurement and monitoring of fidelity to evidence-based interventions (EBIs). Fidelity has been identified as a quality indicator of mental health care (Institute of Medicine, 2015; McLeod et al., 2013), and providing feedback has been associated with practitioners' increased use of fidelity-consistent techniques and decreased use of fidelity-inconsistent techniques (Boxmeyer et al., 2008; Caron & Dozier, 2019; Lochman et al., 2009). Fidelity also is conceptualized as an implementation outcome when clearly defined, reflecting a valid performance standard, and associated with program outcomes (Landsverk et al., 2012).

Despite the importance of fidelity, its measurement is challenging. Particularly elusive are measurement methods that are both efficient (i.e., feasible) and effective (i.e., rigorously evaluated, psychometrically sound; Schoenwald et al., 2011). Observational methods have been a gold standard (Stirman, 2020), but their feasibility in community mental health settings is presumed limited due to resource constraints (Hogue et al., 2021). Until recently, information about fidelity measurement appeared primarily in published treatment efficacy or effectiveness studies and lacked details needed to estimate the associated resource requirements (Schoenwald & Garland, 2013). This knowledge gap will begin to be remedied by the results of a randomized trial (Beidas et al., 2016), which compared the accuracy and cost-effectiveness of observational and other methods to measure fidelity to cognitive behavioral therapy (CBT) for children in community mental health settings.

It is also necessary to evaluate methods to improve the efficiency of observational measurement—but without compromising its effectiveness. One approach is to have clinicians provide self-ratings, which could bypass the need for trained coders. However, for clinician self-ratings to be accurate and reliable, training may be required, as the concordance of therapist and observer ratings of EBI components has typically been low (Caron, et al., 2020). In a recent study, an innovative online training system was used to enable community-based therapists to practice observational coding of family therapy techniques in mock video vignettes (Hogue et al., 2021). There were promising training effects on the reliability of therapist ratings—but less so for accuracy—which led the investigators to conclude that stronger training effects would be needed for a meaningful impact on EBI implementation.

Another way to improve efficiency is to retain observational raters but shorten the duration of the observation. The effectiveness of this strategy can be evaluated by comparing scores based on observing partial versus full sessions. Two studies of CBT sessions have made such comparisons (Weck et al., 2011, 2014). In the first, comparing middle segments and whole sessions, coder judgments were comparable for global adherence and competence to CBT, but inter-rater reliability was higher for specific aspects of adherence and competence when based on entire sessions. In the second, segment and complete session data were compared for treatment integrity ratings from three studies of CBT, and the results supported the adequacy of segments. However, the results of a recent study were less promising: segment coding had lower inter-rater reliability than full session coding of fidelity to a family-focused EBI (Smith et al., 2019).

The present study evaluated the segment coding approach in observational measurement of supervisors' fidelity to an EBI in community practice settings. Clinical supervision is a

quality assurance method commonplace in community mental health settings serving children (Bickman, 2000; Schoenwald et al., 2008). The potential reach of supervisor effects (i.e., a single supervisor can affect multiple therapists, each of whom serves multiple clients) renders supervision a potentially efficient leverage point for scaling and sustaining EBI implementation. Recent studies have begun to evaluate the nature of supervision practices and associated strategies for training and sustaining supervisor implementation in routine care settings (Bearman et al., 2017; Dorsey et al., 2018). Given the effectiveness of A&F in changing the behavior of health care professionals, its effect was evaluated on supervisor fidelity to an EBT supervision protocol, which was measured using an observational coding system (Chapman et al., 2022). The results were promising: The coding system performed well and monthly web-based feedback demonstrated maintenance or improvement of supervisor adherence (and to a lesser extent competence.) For future use of the system, the primary limitation was coding time, which required the full supervision session time ($M = 49.2$ min, $SD = 18.6$) and an estimated 10 additional minutes. The current evaluation uses the same data to evaluate the accuracy of observational coding when based on partial, rather than full, observations of supervision sessions.

Method

Parent Study

The parent study is detailed in Chapman et al. (2022). The study was a prospective, two-arm randomized controlled trial with 60 Multisystemic Therapy® (MST; Henggeler et al., 2009) supervisors randomized to one of two conditions: Supervisor Audit (SA) or Supervisor Audit and Feedback (SAF). The primary outcomes were supervisor Adherence to, and Competence of following, the MST supervision protocol. Adherence and Competence were measured by coding audio-recorded group supervision sessions with the Supervisor Observational Coding System (SOCS; developed and evaluated in the parent study). Supervisors uploaded weekly recordings for 7 months, from which, one session per month (randomly selected) was rated by one of four trained observational coders. Sessions were structured with a series of individual case discussions between the supervisor and team therapists, and according to the SOCS, the first six case discussions were rated. Of 60 enrolled supervisors, 57 (95%) provided at least one session for rating, and 49 (86%) had complete data. A total of 374 sessions were rated, with 165 double-coded (44%). Supervisors in the SAF condition received a series of six, monthly web-based feedback reports. All study procedures, including supervisor informed consent, were approved by the Institutional Review Board of the Oregon Social Learning Center.

Present Study

The present study considered whether ratings from partial sessions would provide similar conclusions to ratings from complete sessions. Two types of partial session observations were considered: (1) a randomly selected *15-min segment* and (2) the session's *first case discussion*. For the 15-min segment, one session per supervisor was randomly selected, the total duration was computed, and a continuous 15-min segment was randomly selected for rating. On average, the 15-min segment covered 36% of the full session duration ($SD = .18$). All segments were rated by two coders. For the first case discussion, the average duration was 11.7 min ($SD = 7.7$), which covered 24% of the full session ($SD = .14$). Ratings of the first case discussion were available for all sessions in the parent study.

Measurement

The SOCS was developed and evaluated in the parent study using methods based on Item Response Theory. There are three theoretical domains: Analytic Process (AP; 10 components; e.g., "Identify advances and barriers"), Principles (P; 9 components; e.g., "Positive and strength-focused"), and Structure and Process (SP; 12 components; e.g., "Identifies top clinical concern"). For AP and P, each component was rated for Adherence, indicating whether it was observed (0 = *No*, 1 = *Yes*). Components that were observed also were rated for Competence, indicating the quality of delivery (1 = *Low*, 2 = *Moderate*, 3 = *High*). Components in the SP domain, which were always applicable, were only rated for Competence. Psychometric performance of the SOCS is detailed in Chapman et al. (2022). Within the AP, P, and SP domains, there was no indication of further dimensionality, and for Competence ratings, the three-point scale performed as intended. From the full sample, the reliability of session-level scores (estimated from multilevel models with cases, sessions, and supervisors) ranged from .27 to .76 for Adherence and .64 to .95 for Competence. For 15-min segments, Rasch separation reliability (ranging from 0 to 1 and generally more conservative than traditional reliability estimates) ranged from .47 to .51 for Adherence and .31 to .61 for Competence. For first case discussions, it ranged from .55 to .66 for Adherence and .48 to .57 for Competence. Inter-rater agreement was based on absolute agreement, and for the full sample, ranged from 87% to 89% for Adherence (15-min segment: 81%–84%; first case discussion: 75%) and 53%–63% for Competence (15-min segment: 44%–65%; first case discussion: 43%–58%). For the present study, SOCS scores were computed as raw averages across all available coders.

Data Analysis Strategy

The analyses compared supervisor fidelity scores based on observation of *partial* sessions—15-min segments and first case discussions—to scores based on *full* sessions. In all models, the predictor was the partial session score, and the outcome was the full session score. Models evaluating 15-min segments had one session per supervisor and were performed as OLS regressions in SPSS. For this sub-

Table 1. Results of OLS Regression Models Evaluating the Association Between Scores From Partial Sessions and Scores From Full Sessions With One Session Sampled per Supervisor.

	15-Minute Segment Predictor				First Case Discussion Predictor			
	Est.	*SE*	*p*	95% CI	Est.	*SE*	*p*	95% CI
Adherence								
Analytic Process (AP)								
Intercept	0.184	0.028	<.001	[0.128, 0.239]	0.190	0.025	<.001	[0.141, 0.240]
Proportion of Session	−0.245	0.057	<.001	[−0.360, −0.131]	−0.236	0.077	.003	[−0.391, −0.082]
Partial Session Score	0.374	0.060	<.001	[0.254, 0.493]	0.413	0.060	<.001	[0.292, 0.533]
Principles (P)								
Intercept	0.164	0.033	<.001	[0.097, 0.231]	0.168	0.029	<.001	[0.109, 0.227]
Proportion of Session	−0.243	0.068	.001	[−0.391, −0.116]	−0.205	0.095	.036	[−0.396, −0.014]
Partial Session Score	0.471	0.072	<.001	[0.327, 0.615]	0.490	0.066	<.001	[0.357, 0.622]
Competence								
Analytic Process (AP)								
Intercept	1.977	0.040	<.001	[1.896, 2.058]	2.003	0.035	<.001	[1.933, 2.072]
Proportion of Session	0.216	0.258	.405	[−0.301, 0.733]	−0.470	0.277	.096	[−1.025, 0.086]
Partial Session Score	0.416	0.118	.001	[0.180, 0.653]	0.456	0.086	<.001	[0.284, 0.629]
Principles (P)								
Intercept	1.972	0.036	<.001	[1.900, 2.044]	1.977	0.031	<.001	[1.915, 2.040]
Proportion of Session	0.042	0.231	.855	[−0.421, 0.506]	−0.626	0.241	.012	[−1.109, −0.143]
Partial Session Score	0.463	0.113	<.001	[0.236, 0.689]	0.464	0.076	<.001	[0.311, 0.617]
Structure & Process (SP)								
Intercept	2.053	0.035	<.001	[1.983, 2.122]	2.053	0.022	<.001	[2.008, 2.098]
Proportion of Session	−0.149	0.220	.501	[−0.590, 0.292]	−0.447	0.175	.013	[−0.797, −0.097]
Partial Session Score	0.566	0.145	<.001	[0.276, 0.856]	0.673	0.064	<.001	[0.545, 0.801]

Note. The sample included one randomly selected session per supervisor ($N = 56$). Two types of partial session predictors were evaluated: The 15-min segment predictor was the Adherence or Competence score for a randomly selected 15-min segment of the session, and the first case discussion predictor was the Adherence or Competence score from the first case discussion of the session. Each model controlled for the proportion of the full session duration covered by the partial session score. The control was grand mean centered, with the resulting intercept reflecting an average proportion.

sample, the first case discussion ($N = 56$) was also evaluated. Additionally, the first case discussion was available for all sessions coded in the parent study. This led to a two-level nested data structure with monthly measurements (level-1; $N = 374$) nested within supervisors (level-2; $N = 57$), modeled as mixed-effects regression models in HLM (Raudenbush et al., 2019). For both data structures, the focal predictor was the partial session score (uncentered for Adherence, grand mean centered for Competence). In the mixed-effects models, the partial session score was time-varying. Finally, sessions were of variable duration; therefore, all models were controlled for the proportion of the session covered by the partial observation (grand mean centered).

Results

One Session per Supervisor

15-Minute Segment

Results are reported in the left section of Table 1 and illustrated in Figures 1 and 2. In all cases, there was a positive and statistically significant association between

Figure 1. Partial session score predicting full session score for adherence (AP and P).
Note. Panel A: Analytic Process Adherence. Panel B: Principles Adherence. The figure illustrates results across the two sets of analyses. The lines for 15-Minute Segment and First Case (Partial Sample) reflect the 15-min segment sample with $N = 56$ supervisors, and the line for First Case (Full Sample) reflects the full sample with a maximum of seven repeated measurements ($N = 374$) nested within supervisors ($N = 57$). The reference line represents a correlation of 1.0, that is, an identical conclusion about Adherence based on the full session score and partial session score. The models controlled for the proportion of the full session duration covered by the partial session observation, with the plotted values reflecting an average proportion.

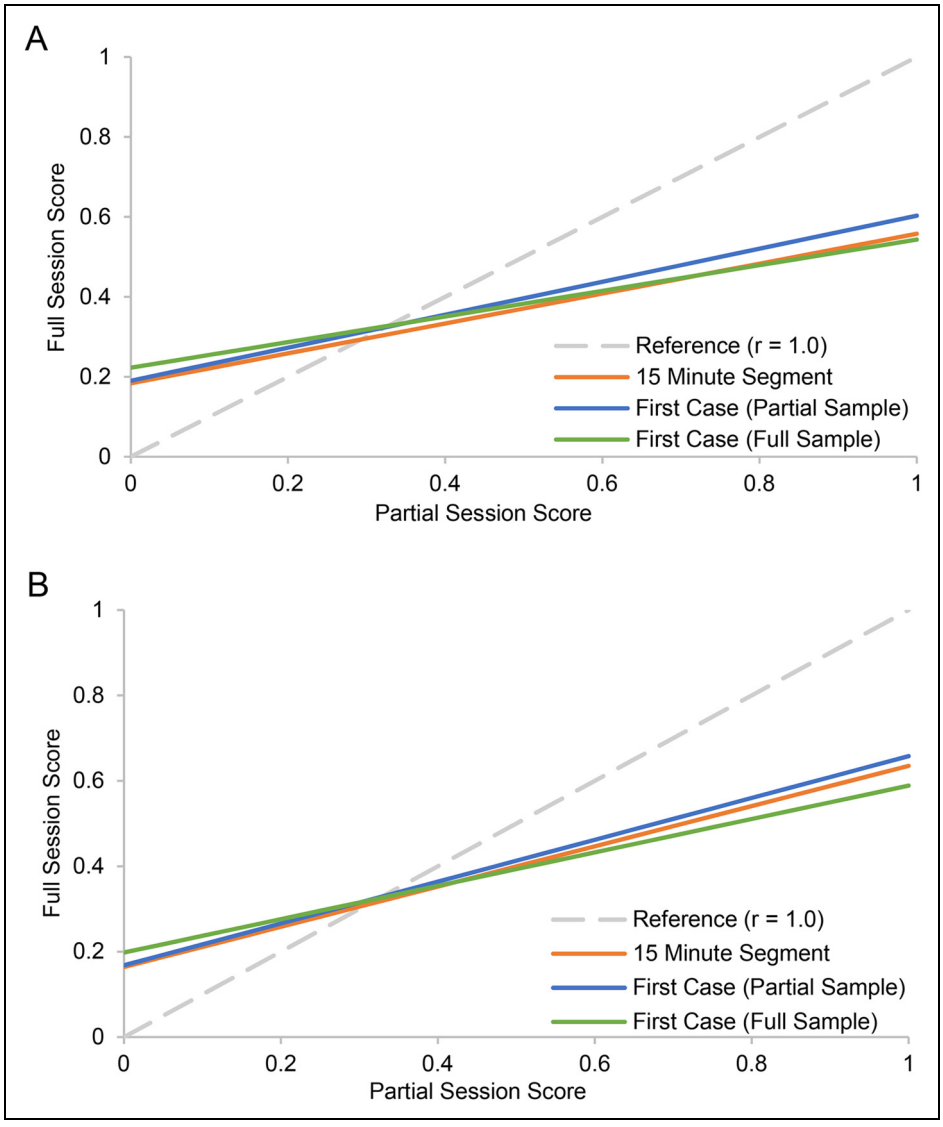
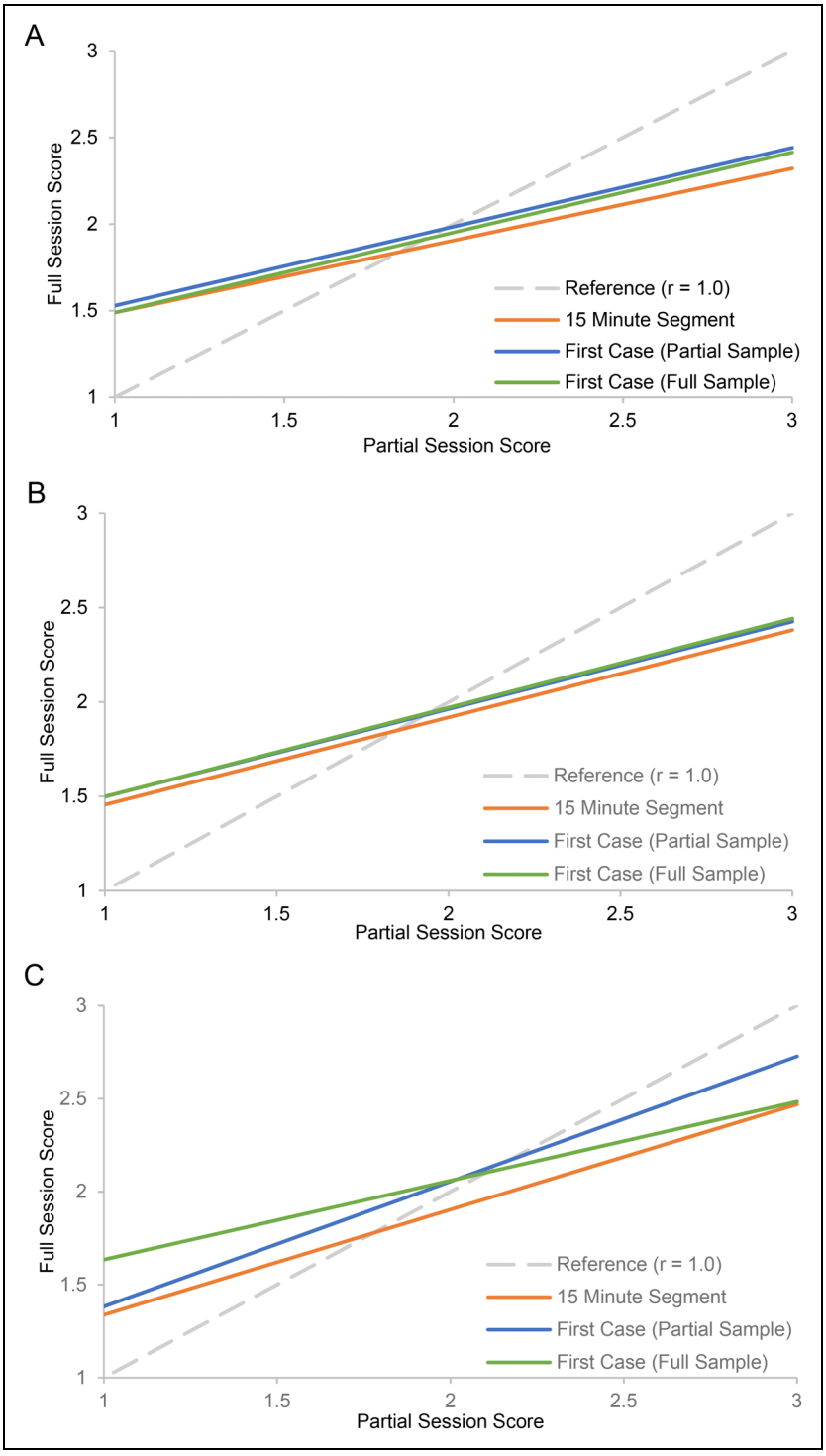


Figure 2. Partial session score predicting full session score for competence (AP, P, and SP).
Note. Panel A: Analytic Process Competence. Panel B: Principles Competence. Panel C: Structure and Process Competence. The figure illustrates results across the two sets of analyses. The lines for 15-Minute Segment and First Case (Partial Sample) reflect the 15-min segment sample with $N = 56$ supervisors, and the line for First Case (Full Sample) reflects the full sample with a maximum of seven repeated measurements ($N = 374$) nested within supervisors ($N = 57$). The reference line represents a correlation of 1.0, that is, an identical conclusion about Competence based on the full session score and partial session score. The models controlled for the proportion of the full session duration covered by the partial session observation, with the plotted values reflecting an average proportion.



supervisor Adherence (AP, P) and Competence (AP, P, SP) scores based on 15-min segments and scores based on full sessions. For Adherence, if the score from a 15-min observation increased by one-point, the score for the corresponding full-session was expected to increase by 0.18 (AP) and 0.16 (P) points. For Competence, a one-point increase for the 15-min segment was associated with full-session increases of 0.42 (AP), 0.46 (P), and

Table 2. Results of Mixed-Effects Regression Models Evaluating the Association Between Scores From the First Case Discussion and Scores From the Full Session in the Complete Sample of Data.

	Fixed Effect Estimates			
	Est.	*SE*	*p*	95% CI
Adherence				
Analytic Process (AP)				
Intercept	0.223	0.012	<.001	[0.199, 0.247]
Proportion of Session	−0.141	0.028	<.001	[−0.196, −0.086]
First Case Discussion	0.321	0.022	<.001	[0.278, 0.364]
Principles (P)				
Intercept	0.198	0.011	<.001	[0.176, 0.220]
Proportion of Session	−0.089	0.027	.001	[−0.142, −0.036]
First Case Discussion	0.391	0.024	<.001	[0.344, 0.438]
Competence				
Analytic Process (AP)				
Intercept	1.969	0.014	<.001	[1.942, 1.996]
Proportion of Session	−0.416	0.095	<.001	[−0.602, −0.230]
First Case Discussion	0.462	0.027	<.001	[0.409, 0.515]
Principles (P)				
Intercept	1.985	0.013	<.001	[1.960, 2.010]
Proportion of Session	−0.427	0.083	<.001	[−0.590, −0.264]
First Case Discussion	0.472	0.026	<.001	[0.421, 0.523]
Structure & Process (SP)				
Intercept	2.058	0.019	<.001	[2.021, 2.095]
Proportion of Session	−0.274	0.064	<.001	[−0.399, −0.149]
First Case Discussion	0.425	0.028	<.001	[0.370, 0.480]
	Variance Component Estimates			
	Est.	*SD*	*p*	
Adherence				
Analytic Process (AP)				
Session	0.004	0.066		
Supervisor	0.002	0.047	<.001	
Principles (P)				
Session	0.004	0.063		
Supervisor	0.001	0.034	<.001	
Competence				
Analytic Process (AP)				
Session	0.052	0.227		
Supervisor	0.002	0.050	.060	
Principles (P)				
Session	0.045	0.212		
Supervisor	0.003	0.050	.048	
Structure & Process (SP)				
Session	0.023	0.151		
Supervisor	0.018	0.134	<.001	

Note. The two-level mixed-effects regression model included a maximum of seven monthly measurements (level-1; $N = 374$) nested within supervisors (level-2; $N = 57$). The first case discussion predictor (time-varying) was the Adherence or Competence score from the first case discussion of each session. Each model controlled for the proportion of the full session duration (time-varying) covered by the 15-min segment. The control was grand mean centered, with the resulting intercept reflecting an average proportion.

0.57 (SP) points. Additionally, the control variable (i.e., proportion of full session covered by 15-min segment) was negative and statistically significant for Adherence but not Competence. This indicates that shorter sessions had lower full-session Adherence scores.

First Case Discussion

Results are reported in the right section of Table 1 and illustrated in Figures 1 and 2. The pattern was consistent with that described for 15-min segments. Specifically, Adherence (AP, P) and Competence (AP, P, SP) scores based on the first case discussion were positively and significantly associated with full session scores. For Adherence, if the score for the first case discussion increased by one-point, the score for the corresponding full-session was expected to increase by 0.19 (AP) and 0.17 (P) points. For Competence, a one-point increase in the first case discussion was associated with full-session increases of 0.46 (AP), 0.46 (P), to 0.67 (SP) points. The control variable was negative and statistically significant for Adherence, and for Competence, it was negative and significant for P and SP. Thus, when the first case discussion covered a larger proportion of the session, full-session scores were lower.

Multiple Sessions per Supervisor

First Case Discussion

With multiple sessions per supervisor, mixed-effects regression models were performed with repeated sessions (level-1) nested within supervisors (level-2), and the score for the first case discussion of each session was entered as a predictor of the full session score. Results are reported in Table 2 and illustrated in Figures 1 and 2. There was no evidence of significant variability across supervisors in the association between scores from the first case discussion and full-session. Consistent with the prior results, Adherence (AP, P) and Competence (AP, P, SP) scores for the first case discussion were positively and significantly associated with scores for the full session. For Adherence, if the score for the first case discussion increased by one-point, the score for the corresponding full-session was expected to increase by 0.22 (AP) and 0.20 (P) points. For Competence, a one-point increase in the first case discussion was associated with full-session increases of 0.46 (AP), 0.47 (P), or 0.43 (SP) points. The control variable was negative and statistically significant in all models, indicating that full-session scores were lower when the first case discussion covered a larger portion of the session.

Discussion

This study, leveraging data from a randomized trial of an A&F intervention for supervisor Adherence and Competence (Chapman et al., 2022), evaluated a more pragmatic approach to observational coding. Specifically, two types of partial session observations—random 15-min segments and first case discussions—were compared to observations from full sessions. Across all models (and adjusting for length of observation), Adherence and Competence scores from partial and full sessions were positively and significantly associated. However, partial session observations did not precisely replicate scores from full sessions. As evidenced in Figures 1 and 2, if the full observation had a low level of Adherence or Competence, the partial observation tended to underestimate it, and if the full observation had a high level, the partial observation tended to overestimate it. This likely reflects variability in Adherence and Competence during the course of sessions; that is, more extreme levels (low or high) likely would not be maintained throughout the session. Thus, in the present supervision context, partial session scores were most accurate when the level of Adherence or Competence was moderate. Future investigations should evaluate the relative accuracy of Adherence versus Competence ratings. In terms of feasibility, randomly selected 15-min segments required specific calculations for identification, and as such, coding the first case discussion was more efficient. However, the first case discussion is vulnerable to validity threats, whether due to supervisors' awareness of the coding plan or systematic characteristics of the case (e.g., supervisor attention/focus, prioritization of challenging cases).

The study had two main limitations. First, for coding full versus partial sessions, resource requirements have not been formally evaluated and compared. This could be important information for determining the feasibility of use in community practice settings. Second, in the parent study, only full-session scores informed feedback to supervisors, and as such, it is unknown whether A&F effects would be retained with feedback based on partial observations.

Conclusion

To support effective supervision practices in community settings, efficient and effective measurement methods—methods that are feasible and psychometrically sound—are needed for monitoring fidelity to EBIs. The present results suggest that efficient observational measurement can be achieved while retaining a general level of measurement effectiveness. The results enhance prior findings supporting partial session coding for EBI fidelity, and they extend the body of research to supervision practices. Building from this, the use of partial session coding to inform quality assurance and improvement efforts, such as A&F interventions, will require empirical evaluation. However, this pragmatic method provides a promising avenue to support the implementation of effective supervision practices.

Acknowledgments

The authors wish to thank 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

Declaration of Conflicting Interests

Sonja K. Schoenwald is a founder and shareholder of MST Services, 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. She is also a Co-Founding Editor in Chief of *Implementation Research and Practice* and took no part in the editorial process for this manuscript. All decisions about the manuscript were made by another editor.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institute of Mental Health (grant number R21/R33MH097000).

ORCID iDs

Jason E. Chapman https://orcid.org/0000-0002-9396-5877
Zoe M. Alley https://orcid.org/0000-0003-0583-8382
Sonja K. Schoenwald https://orcid.org/0000-0002-0560-8440

References

Beidas, R. S., Maclean, J. C., Fishman, J., Dorsey, S., Schoenwald, S. K., Mandell, D. S., Shea, J. A., McLeod, B. D., French, M. T., Hogue, A., Adams, D. R., Lieberman, A., Becker-Haimes, M., & Marcus, S. C. (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: Project FACTS study protocol. *BMC Psychiatry*, *16*(1), 323. https://doi.org/10.1186/s12888-016-1034-z

Bearman, S. K., Schneiderman, R. L., & Zoloth, E. (2017). Building and evidence base for effective supervision practices: An analogue experiment of supervision to increase EBT fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*(2), 293–307. https://doi.org/10.1007/s10488-016-0723-8

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

Boxmeyer, C. L., Lochman, J. E., Powell, N. R., Windle, M., & Wells, K. (2008). School counselors' implementation of coping power in a dissemination field trial: Delineating the range of flexibility within fidelity. *Report on Emotional and Behavioral Disorders in Youth*, *8*(4), 79–95.

Caron, E. B., & 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*(4), 445–457. https://doi.org/10.1007/s10488-019-00924-3

Caron, E. B., Muggeo, M. A., Souer, H. R., Pella, J. E., & Ginsburg, G. S. (2020). Concordance between clinician, supervisor and observer ratings of therapeutic competence in CBT and treatment as usual: Does clinician competence or supervisor session observation improve agreement? *Behavioural and Cognitive Psychotherapy*, *48*(3), 350–363. https://doi.org/10.1017/S1352465819000699

Chapman, J. E., Schoenwald, S. K., Sheidow, A. J., & Cunningham, P. B. (2022). Performance of a Supervisor Observational Coding System and an Audit and Feedback Intervention. Administration and Policy in Mental Health and Mental Health Services Research, *49*(4), 670–693. https://doi.org/10.1007/s10488-022-01191-5

Colquhoun, H. L., Carroll, K., Eva, K. W., Grimshaw, J. G., Ivers, N., Michie, S., & Brehaut, J. C. (2021). Informing the research agenda for optimizing audit and feedback interventions: Results of a prioritization exercise. *BMC Medical Research Methodology*, *21*(1), 20. https://doi.org/10.1186/s12874-020-01195-5

Dorsey, S., Kerns, S. E. U., Lucid, L., Pullmann, M. D., Harrison, J. P., 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. https://doi.org/10.1186/s13012-017-0708-3

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (2009). *Multisystemic therapy for antisocial behavior in children and adolescents* (2nd ed.). Guilford Press.

Hogue, A., Porter, N., Bobek, M., MacLearn, A., Bruynesteyn, L., Jensen-Doss, A., Dauber, S., & Henderson, C. E. (2021). Online training of community therapists in observational coding of family therapy techniques: Reliability and accuracy. *Administration and Policy in Mental Health and Mental Health Services Research*, *49*(1), 131–151. https://doi.org/10.1007/s10488-021-01152-4

Institute of Medicine. (2015). *Psychosocial interventions for mental and substance use disorders: A framework for establishing evidence-based standards*. National Academies Press.

Landsverk, J. A., Brown, C. H., Chamberlain, P., Palinkas, L. A., Ogihara, M., Czaja, S., Goldhaber-Fiebert, J. D., Rolls Reutz, J. A., & Horwitz, S. M. (2012). Design and analysis in dissemination and implementation research. In Brownson, R. C., Colditz, G. A., & Proctor, E. K. (Eds.), *Dissemination and Implementation Research in Health: Translating Research to Practice* (pp. 225–260). Oxford University Press.

Lochman, J. E., 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*(3), 397–409. https://doi.org/10.1037/a0014514

McLeod, B. D., Southam-Gerow, M. A., Tully, C. B., Rodriguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice*, *20*(1), 14–32. https://doi.org/10.1111/cpcs.12020

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2019). HLM 8: Hierarchical linear & nonlinear modeling (version 8.0.2010.18) [Computer software and manual]. Scientific Software International. https://www.ssicentral.com

Schoenwald, S. K., Chapman, J. E., Kelleher, K., Hoagwood, K. E., Landsverk, J., Stevens, J., Glisson, C., & Rolls-Reutz, J., & 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*(1–2), 84–97. https://doi.org/10.1007/s10488-007-0147

Schoenwald, S. K., & Garland, A. F. (2013). A review of treatment adherence measurement methods. Psychological Assessment, *25*(1), 146–156. https://doi.org/10.1037/a0029715

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(1), 32–43. https://doi.org/10.1007/s10488-010-0321-0

Smith, J. D., Rudo-Stern, J., Dishion, R. J., Stormshak, E. A., Montag, S., & Brown, K., Ramos, K., Shaw, D. S., & Wilson, M. N. (2019). Effectiveness and efficiency of observationally assessing fidelity to a family-centered child intervention: A quasi-experimental study. *Journal of Clinical Child and Adolescent Psychology*, *48*(1), 16–28. https://doi.org/10.1080/15374416.2018.1561295

Stirman, S. W. (2020). Commentary: Challenges and opportunities in the assessment of fidelity and related constructs. *Administration and Policy in Mental Health and Mental Health Services Research*, *47*(6), 932–934. https://doi.org/10.1007/s10488-020-01069-4

Weck, F., Bohn, C., Ginzburg, D., & Stangier, U. (2011). Assessment of adherence and competence in cognitive therapy: Comparing session segments with entire sessions. *Psychotherapy Research*, *21*(6), 658–669. https://doi.org/10.1080/10503307.2011.602751

Weck, F., Grikscheit, F., Höfling, V., & Stangier, U. C (2014). Assessing treatment integrity in cognitive-behavioral therapy: Comparing session segments with entire sessions. *Behavioral Therapy*: *45* (4), 541–552. https://doi.org/10.1016/j.beth.2014.03.003