Title

Efficient Use of Clinical Decision Supports: An Evaluation of Change Over Time in the Context of Clinical Supervision

Author

Knudsen, Kendra

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Efficient Use of Clinical Decision Supports:

An Evaluation of Change Over Time in the Context of Clinical Supervision

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Kendra Sue Knudsen

2024

ABSTRACT OF THE DISSERTATION

Efficient Use of Clinical Decision Supports:

An Evaluation of Change Over Time in the Context of Clinical Supervision

by

Kendra Sue Knudsen

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2024

Professor Bruce F. Chorpita, Chair

Recent research highlights a growing demand for youth mental health services (Barican et al., 2022; Kazdin, 2019; USPSTF, 2022), prompting the need to enhance mental health workforce capacity. Improving workforce capacity entails strengthening critical decision-making activities, including considering client problems, prioritizing them, and selecting the most suitable practices to address them. Clinical supervision, involving dyads of qualified mental health professionals ("supervisors") and direct service providers ("supervisees"), aims to improve these activities (Proctor, 1986; Milne, 2007). Challenges include time constraints, varying competency activity levels, and difficulty in incorporating new scientific findings, compounded by high turnover rates (Bernstein et al., 2015; Brabson et al., 2020; Chorpita et al., 2021; Collatz & Wetterling, 2012; Dorsey et al., 2017; Powell & York, 1992; Simon & Greenberger, 1971). Integrating decision support systems into clinical supervision could address these challenges, promoting use of evidence and ensuring sustained skill retention among supervisory dyads (Bjork & Bjork, 2020).

Within the context of a decision-support system integrated within clinical supervision, this dissertation investigated the reliability of quality, effort, and efficiency metrics, and then examined the

associations between ordinal repetition of activities and passage of time with those quality and effort metrics. As such, it explored whether time or repetition is associated with improvement, deterioration, or no change in these metrics.

The study analyzes existing data from a multi-site randomized implementation trial aimed at promoting the use of evidence-based methods for engaging youth and families in treatment. We audio recorded and transcribed supervision events in which mental health workers discussed cases at-risk for poor treatment engagement. For part one, 26 supervisees and 17 supervisors discussed 30 cases; for part two, 48 supervisees and 16 supervisors, trained and using a decision-support system, discussed 118 cases.

Observational coders rated efficiency and the extensiveness of decision-making activities using a subset of the ACE-BOCS coding system (Chorpita et al., 2018). Efficiency was rated holistically for each event on a 5-point scale, from presence of extensive discussions on unnecessary topics (1) to swift and organized decision-making and planning (5). Quality was evaluated using a dichotomous scale, based on whether each activity met sufficient quality criteria, primarily indicating the presence of the activity. Effort was measured by the total number of words spoken for each activity. Two overall effort scores were calculated based on the total words spoken and duration of the entire event. The total number of supervisory events per supervisory dyad was an indicator of repetition of supervisory activities, and the total weeks since training in the decision-support system measured the passage of time.

To assess interrater reliability across all coders, we used Fleiss' kappa (κ) for the four dichotomous quality metrics and *ICC*s (model [2,1], consistency) for the ordinal efficiency metric. To examine possible change in outcomes, we used mixed effects regression models, examining three hierarchical levels: cases nested within supervisees nested within supervisors. Thus, supervisors were the main level of analysis. We assessed the impact of each level on results and simplified the model if it didn't improve it. To manage skewed data with quality and effort measures having excess zeros, we implemented corrections like the Firth logistic regression and employed specialized models such as the Hurdle model, respectively. These strategies helped mitigate bias and stabilize parameter estimates.

Interrater reliability estimates showed that coders consistently rated both the decision-making activities and overall efficiency reliably. A strong positive correlation confirmed the initial validity of the effort measure. Findings revealed changes in efficiency, the presence of quality, and the likelihood of putting in effort as dyads moved through each level of supervision for their cases (for example, from the first supervision event type to the second and then to the third type). Increasing repetition of supervision events or time within each supervision stage did not predict whether the dyads improved in these outcomes.

This study underscores the sustainability of quality, effort, and efficiency across repeated supervision events within different supervision types and over time. It also identifies areas for further investigation, including the need for more nuanced and robust measures of quality and effort. Future research should address these issues and explore alternative assessment methods to gain a deeper understanding of workforce learning. This understanding will inform strategies aimed at maximizing workforce capacity to meet the growing demand for high-quality youth mental health services.

This dissertation of Kendra Sue Knudsen is approved.

Kimberly D. Becker

Robert M. Bilder

Craig K. Enders

Bruce F. Chorpita, Committee Chair

University of California, Los Angeles

2024

Contents

Figures and Tables

Vita

EDUCATION

2023	UCLA Semel Institute
	Pre-doctoral Internship
2019	University of California, Los Angeles
	Master of Arts in Psychology
2013	University of California, Los Angeles
	Bachelor of Arts in Psychobiology, Departmental Honors

SELECTED AWARDS

2021, 2023	Charles F. Scott Fellowship
2019, 2020	UCLA Graduate Summer Research Mentorship Fellowship
2020	UCLA Affiliates Fellowship
2019	Graduate Research Mentorship Fellowship

SELECTED PUBLICATIONS

Knudsen, K.S., Becker, K.D., Guan, K., Gellatly, R., Patel, V., Malik, K., Boustani, M., Mathur, S., & Chorpita, B. F. (2021). A pilot study to evaluate feasibility and acceptability of training mental health workers in India to select case-specific intervention procedures within a dynamic modular treatment designed for a low-resource setting. *Journal of Evaluation in Clinical Practice, 28*(4), 531-541.

Gellatly, R., Knudsen, K.S., Boustani, M., Michelson, D., Malik, K., Mathur, S., Pooja, N., Patel., V., and Chorpita, B. F. A Qualitative Analysis of Collaborative Efforts to Build a Multi-Problem, School-Based Intervention for Common Adolescent Mental Health Difficulties in India (2022). *Frontiers in Psychiatry*.

Anderson, A., Japardi, K., Knudsen, K.S., Bookheimer S., Ghahremani, D.G., and Bilder, R. (2021). Big-C creativity in artists and scientists is associated with more random global but less random local patterns of fMRI functional connectivity. *Psychology of Aesthetics, Creativity, and the Arts*, 1- 11.

Knudsen, K. S., Bookheimer, S., and Bilder, R. M. (2019). Is psychopathology elevated in Big-C visual artists and scientists? *Journal of Abnormal Psychology, 128*(4), 273.

Introduction

Recent data reveal that more than 13% of youth report mental health needs (Barican et al., 2022), a figure expected to surge with new recommendations advocating for mental health screenings among children aged 8 and above (USPSTF, 2022). Furthermore, the COVID-19 pandemic has laid bare the inadequacies of existing mental health systems, accentuating the urgency for refined mental health service delivery methods (Kazdin, 2019). Given the current landscape, there exists a pressing need to cultivate mental health worker capacity. Workforce capacity refers to the collective ability of mental health workers to meet the diverse and increasing needs of individuals and families seeking mental health services. In essence, it encompasses the skills, knowledge, resources, and organizational structures required to deliver effective and responsive care to a broad range of clients.

The challenge of maintaining and enhancing workforce capacity within mental health systems is multifaceted. A recent study (Chorpita et al., 2021) compared the effectiveness of various implementation strategies in building a prepared workforce by analyzing archival data from a clinical trial. The analysis revealed that nearly all implementation strategies tested were only able to reach approximately half of the population in need within 60 days, underscoring the difficulty in rapidly developing a wide and diverse range of competencies to strengthen workforce capacity. Compounding this challenge is the fact that a limited number of mental health workers are tasked with mastering numerous competencies to effectively serve the population. Moreover, the diverse baseline competency levels and learning rates among incoming professionals exacerbate the complexities of training (Simon & Greenberger, 1971). Additionally, the constant influx of new scientific findings adds to the challenge, creating what is commonly known as the "wealth of information problem" (Bernstein et al., 2015; Collatz & Wetterling, 2012). As mental health workers strive to incorporate these findings into their practice, they face limitations in their capacity to act on the vast amount of available information swiftly and effectively. Furthermore, workforce turnover poses a significant obstacle, with estimates in public mental health systems ranging from 30 to 60 percent, and up to 100 percent within a 4-year period (Brabson et al.,

2020). Turnover without adequate replacement reduces workforce capacity, whereas turnover with replacement risks eroding baseline competency levels, as new entrants often lack the skills of their predecessors (Powell & York, 1992). Thus, in this challenging environment, mental health service systems must continually invest in ongoing training initiatives to maintain therapeutic impact within communities.

To address some of these challenges to workforce capacity, some researchers have called for increasing the size of the workforce through paraprofessional extensions, such as task sharing or shifting, to improve the efficiency and thus the work output of human providers. Task sharing/shifting makes use of paraprofessionals or professionals from non-mental health domains (Patel et al., 2010; Rotheram-Fuller et al., 2017). Although this strategy may increase the workforce's overall effort (input), one limitation to increasing the workforce's size is that this typically also raises the burden of ensuring workforce's overall quality in skill level (output). Some researchers have also advocated for machines to improve the efficiency of human providers, such as by delivering services that might not require human providers (Muñoz et al., 2016). Compared with human-delivered services, however, machine interventions have so far shown widely variable and problematic rates of disengagement, attrition, and reduced acceptability, complicating their potential quality (Andrews et al., 2018; Borghouts et al., 2021; Kaltenthaler et al., 2008). Other models have considered geographic distribution or mapping of the workforce as ways of increasing productive capacity (Salvador-Carulla et al., 2015); yet even large-scale, coordinated efforts cannot fully remedy the substantial gap between available mental health providers and people with mental health needs, still rendering the need for increasing workforce capacity.

Enhancing workforce capacity involves increasing workforce competency, which is defined as the array of skills or actions mastered by individuals at a specific time. For instance, competencies may include adeptness in guiding individuals through exposure procedures to manage anxiety, conducting thorough assessments to identify underlying mental health issues, effectively managing therapeutic alliances to foster trust and collaboration, or efficiently managing clinical administration tasks such as documentation to ensure accurate record-keeping and compliance with regulatory standards, exemplifying

the breadth of skills encompassed within workforce competency. Workforce competency is a multifaceted and multilevel construct (Chorpita & Daleiden, 2014): different systems exhibit varying average slopes for diverse competencies, with individuals within those systems demonstrating slopes that fluctuate around the average competency level of the workforce. These slopes represent the developmental potential of both systems and individual members, indicating the differing rates at which workforce individuals can enhance their performance, often referred to as "absorptive capacity" in management literature (Cohen & Levinthal, 1990). This capacity is inherently finite and can be empirically measured (Knudsen, 2020; Knudsen et al., 2021).

The competencies of problem classification and practice selection are critical to workforce capacity. Problem classification entails accurately identifying and comprehending the specific mental health issues presented by clients, laying the groundwork for tailored treatment plans. Similarly, practice selection involves choosing evidence-based interventions best suited for addressing clients' needs and goals. It is increasingly evident that decision-making regarding problem classification forms the bedrock of all evidence-based practice selection and implementation (Youngstrom et al. 2015, 2017; Youngstrom and Van Meter 2016). Thoroughly assessing client problems heightens the likelihood that the practices employed to address them are evidence-based (Hunsley & Mash, 2020), and discussions regarding client problems and practices correlate with positive supervisory outcomes (Bradley & Becker, 2021). By refining these competencies, mental health workers can augment their ability to deliver personalized, evidence-based care that is attuned to the diverse needs of their clients, ultimately fostering better service outcomes.

These themes of increasing workforce capacity and performance in competencies have been a focus of implementation science for more than 20 years (Glasgow, 1999), yet progress has been limited. Traditionally, tactics to build competencies across these workforces have involved classroom-based teaching, evidence-based training workshops, clinical supervision, and expert consultation (Becker-Haimes, Lushin, et al., 2019; Herschell et al., 2010; Sholomskas et al., 2005). However, these methods have multiple limitations. Classroom-based trainings, which typically involve a combination of (passive)

didactic instruction with active learning through homework and clinical practicum (Becker-Haimes, Okamura, et al., 2019), often overlook the use of evidence-based practices (Weissman et al., 2006). Evidence-based training workshops typically attempt to condense a lot of information into few events (Becker-Haimes, Lushin, et al., 2019) and thus often fail to induce lasting changes in behavior (Weissman, Verdeli et al., 2006; Bertram, Charnin et al., 2015; Scott, Klech et al., 2016; McHugh and Barlow, 2010; Beidas and Kendall, 2010).

Considered as perhaps the most important training experience for developing workforce competencies (Falender & Shafranske, 2004; Stoltenberg, 2005), clinical supervision can vary greatly in quality across contexts (Bailin et al., 2018; Dorsey et al., 2017; Fukui et al., 2014) and is often severely time-limited in school and community settings (Accurso et al., 2011; Dorsey et al., 2017; Lucid et al., 2018). For example, one observational study noted that supervisory teams had on average 30 minutes to an hour each week to cover an average caseload of 31 clients (Dorsey et al., 2017). Ongoing consultation, although effective, is resource intensive (Beidas and Kendall, 2010; Herschell, Kolko et al., 2010). Evidence reveals that ongoing follow-up and evaluation-based guidance – such as through clinical supervision or expert consultation – can better sustain workforce learning, change behaviors, and build MHW's competencies to improve clinical outcomes (Bearman et al., 2013; Beidas et al., 2012; Brookman-Frazee et al., 2012; Edmunds et al., 2013; Rakovshik et al., 2016; Regan et al., 2019; Ruzek et al., 2014). Unfortunately, overall findings on behavioral change through supervision and consultation have been modest (Beidas & Kendall, 2010; Beidas et al., 2019; Frank et al., 2020; Herschell et al., 2010; Monson et al., 2018; Rakovshik & McManus, 2010). Thus, these traditional approaches often prove insufficient, underscoring the imperative for innovative methods to support mental health workforce development.

Research into the fundamental mechanisms of learning (Bjork & Bjork, 2020) underscores the potential for decision support systems coupled with clinical supervision post-training to address the limitations of traditional methods to facilitate enduring retention of knowledge and skills. This science of learning highlights the importance of "desirable difficulties" to enhance long-term information retention.

According to this body of work, proven approaches to creating "desirable difficulties" include active engagement, interleaving different materials, varying learning contexts, and spacing out learning events over time. Within the context of clinical supervision, decision support systems can leverage these desirable difficulties to enhance learning. By employing techniques like filtering, aggregating, and layering information (Simon and Greenberger 1971), decision support systems adeptly manage data and prompt workers to actively devise solutions based on their comprehension of individual client needs. This approach may cultivate competence through fostering skill generalization across various scenarios. The structured framework of clinical supervision further provides spaced learning opportunities, departing from passive review of treatment manuals or the cramming approach common in single-session workshops. Additionally, engaging in dyadic discussions during spaced learning enables workers to reflect on their experiences and refine decision-making strategies over time, contributing significantly to continuous improvement. Thus, the integration of decision support systems within the framework of clinical supervision holds promise in addressing the shortcomings of traditional training methods and fostering enduring retention of knowledge and skills.

Our recent pilot study in India highlights the considerable potential of decision-support systems in supporting baseline mental health worker capacity (Knudsen, Becker et al. 2021). In this study, we examined the preliminary effectiveness of a one-page decision-support strategy to prepare MHWs in low-resource context, India, to classify problems and select practices within a flexible, modular cognitive behavioral intervention (Chorpita et al., 2020). Before the training and without the resource, MHWs classified problems at below chance levels and selected practices at no better than chance levels, compared with decisions made through consensus between psychologists (“judges") with expertise in modular protocols. However, after the training and while using the resource, the MHWs’ rate of agreement with the judge criterion on problem selections increased almost two-fold. This pilot study lends support to the notion that the baseline capacities of mental health workers in problem identification and intervention selection are not fixed but can be rapidly strengthened.

Learning in the context of improving workforce competencies can be evaluated through various lenses, including efficiency, quality, and effort. Efficiency pertains to the ability of mental health workers to achieve desired outcomes with minimal time and effort expended. It involves streamlining processes, optimizing workflows, and maximizing productivity while maintaining high standards of service delivery. Quality, on the other hand, focuses on the effectiveness of mental health services provided. It encompasses factors such as how thoroughly one adheres to evidence-based practices. Quality learning involves continuously refining skills, updating knowledge, and integrating feedback to enhance service provision and ensure positive outcomes for clients. Finally, effort refers to the investment of various resources such as time, energy, and communication in the learning process. This includes the amount of time spent engaged in discussions and the allocation of other resources necessary for skill acquisition. By measuring learning in terms of efficiency, quality, and effort, we can gain a comprehensive understanding of how workforce competencies may develop with practice or the passage of time.

By knowing whether and how specific competencies become efficient (and by implication, consolidated for the learner), we may then train more competencies on top of the first set to improve MHW's capacity. If we are to have an efficient workforce, we need to know how newly trained competencies develop over time -- in terms of their quality output and effort spent. In other words, before we train MHWs in a new competency, we need to know that their current competencies are of sufficient high quality and can be done quickly enough that there is remaining capacity for them to learn the next competency. We currently know little about what the workforce development process looks like if we intend to have developmentally sensitive decision support tools, whose complexity unfolds as the user competency grows.

The Present Study

Thus, the objective of this two-part dissertation was to explore the development of mental health workers in moderately to well-resourced and high-capacity systems. Our focus was on investigating the relationships between the sequential repetition of supervisory activities and the passage of time with

measures of effort, quality, and efficiency within the framework of decision-making support during supervision.

The first part of the study sought to establish reliable and valid measures of supervisory efficiency for problem classification and practice selection. This involved developing and evaluating separate quality and effort scores, along with an efficiency metric, specifically tailored to these competencies. By analyzing these components separately, the hope was to gain a more precise understanding of their development.

The second part of the study adopts a developmental approach to performance. We explored whether time or repetition of supervisory activities was associated with improvement, deterioration, or no change in effort, quality, and efficiency. We used the quality, effort, and efficiency metrics established in the initial study. We analyzed time in two ways: by counting the number of supervision events among supervisory dyads and by tracking the time in weeks since the initial training in the decision support system. This approach helped us determine whether it was repetition, rather than simply the passage of time, that contributed to the establishment of supervisory efficiency in problem classification and practice selection.

To better understand mental health worker learning over time or with repetition of supervision events, we conducted separate analyses of quality, effort, and efficiency. This approach allowed us to determine whether any changes or lack thereof in efficiency over time or with practice could be attributed to reduced supervisory effort on specific activities, enhanced quality in these activities, or a combination of both factors. For instance, if overall efficiency improved, it would be possible that quality improved but effort stayed the same, or quality stayed the same but effort decreased, or both quality improved and effort decreased. Alternatively, it is possible that effort increased at a much higher magnitude when quality improved, thus producing a net overall decrease in efficiency. These configurations could also change across various stages of learning; for example, efficiency could become worse before it improves.

Method

The present dissertation used archival data gathered from a multisite randomized controlled trial of a structured decision support system ("coordinated knowledge system;" CKS) designed to improve clinical decision making in supervision to address mental health treatment engagement challenges. Data collection occurred between August 2017 and May 2020. The trial included multiple school and community mental health clinics that serve youth and caregivers in urban and rural America. Both sites experienced time-limited clinical supervision and high rates of poor engagement in mental health services.

Study Participants

Study participants for part 1 of the study (*N*=43) included 26 direct-service providers and their 17 supervisors who discussed a total of 30 study cases identified as at-risk for poor engagement across both trial conditions, described below. Study participants for study part 2 of the study (*N*=64) involved those in the experimental condition only; this included a total of 48 direct-service providers and their 16 supervisors who discussed a total of 118 study cases identified as at-risk for poor engagement.

The sample of supervisors (*n* = 17) included in part 1 of the study had an average age of 47.10 years (SD = 9.41). Predominantly female (94.11%), they identified as Black/African American (52.94%), White/European American (23.53%), Latina/o/x (17.65%), Asian American (11.76%), and Belizean (5.88%). A majority held master's degrees (94.12%). Supervisors had accrued a mean of 15.47 (SD = 6.76) years of full-time clinical experience following degree completion. The sample of providers (*n* = 26) included in part 1 of the study were about 39.08 years old (SD = 10.14). Most providers identified as female (92.3%), and they identified as Black/African American (50.00%), Latina/o/x (30.77%), White/European American (11.54%), and Asian American (7.69%). The majority obtained master's degrees (96.15%). On average, these providers had participated in clinical work for 7.21 (SD = 6.26) since degree completion.

The sample of supervisors (n = 16) included in part 2 of the study were 42.28 years old on average (SD=10.23), predominantly female (93.75%). They identified as Black/African American (50%), White/European American (25.00%), Latina/o/x (18.75%), Asian American (0%), and multi-racial (6.25%). A majority held master's degrees (75%). Supervisors had accrued a mean of 15.35 years (SD = 7.80) of clinical experience since degree completion. The sample of providers (n = 48) included in part 2 of the study had an average age of 36.70 yeas (SD = 7.78) and predominantly identified as female (91.67%). They identified as Black/African American (41.67%), Latina/o/x (45.83%), White/European American (8.33%), and Asian American (4.17%). Many obtained master's degrees (48%). On average, these providers had participated in clinical work for 6.01 years (SD = 5.06) following degree completion.

Study Conditions

Experimental Group: Coordinated Knowledge System (CKS)

Providers and supervisors in the experimental group underwent a 12-hour, 1.5-day training on a Coordinated Knowledge System (CKS), which featured various decision support resources. These resources included a concise worksheet and eleven detailed "Engagement Guides," which outlined detailed steps of engagement practices. The CKS facilitated problem classification across five discrete dimensions of engagement problems (Becker et al., 2018) and provided several evidence-supported practices to address each problem. The training focused on utilizing the worksheet during supervision to identify engagement problems, select appropriate practices, and assess outcomes. This process, known as the CARE process (Chorpita & Daleiden, 2014), was based on established decision-making principles (Deming, 1989). The training employed active learning methods such as modeling, roleplaying, and group discussions.

Comparison Group

The comparison group received a 30-minute overview of engagement problems and practices, along with a one-page resource listing the names and brief definitions of engagement practices, without a model or detailed steps for their use.

Procedure

Participants in this dissertation were sampled from the Reaching Families Study, a multi-site trial investigating how a coordinated knowledge system affects therapists' use of evidence and client engagement in school-based mental health services. The therapists and supervisors were from the Los Angeles Unified School District in urban Los Angeles, California, or the South Carolina Department of Mental Health in rural South Carolina.

Mental health services administrative staff administered engagement screeners to youth and their caregivers, either online or in person. All caregivers who participated in the youth's treatment were eligible for screening. The surveys were administered about 4-6 weeks after the youth enrolled in mental health services. This timing meant the first recorded supervision events were not the first to occur. Only the study team could access completed online surveys, and paper surveys were also kept confidential from mental health providers.

The research team evaluated the screeners using the "REACH" engagement subscales to identify engagement concerns. A case was eligible if the youth or caregiver scores indicated concerns on at least one subscale. Therapists were notified of eligible cases via a HIPAA-secure email, which included a standard message about the case's eligibility and, for CKS participants, a graphical report of the scores.

During the next treatment session, mental health providers explained the study and obtained informed consent from youth and caregivers in either English or Spanish. Supervisees and supervisors then recorded three supervision events and two therapy events for each eligible and consenting youth or caregiver. The sequence of digitally recorded supervision and therapy sessions in this study can be represented as S1 – T1 – S2 – T2 – S3. These recordings were transcribed, triple-checked for accuracy, translated if needed, and coded using the Action Cycle and Use of Evidence Behavioral Observation Coding System (ACE-BOCS), an observational coding system evaluating evidence use in decision-making (Chorpita et al., 2018). All staff involved were blind to the participants' study conditions.

Study 1 involved supervisors overseeing between two and six therapists ($M = 3.53$, $SD = 1.42$), with each supervisee handling one to four cases ($M = 2.58$, $SD = 0.99$). Overall, supervisors managed

three to thirteen cases (M = 7.94, SD = 3.21) and took part in five to 39 supervision events (M = 23.00, SD = 9.75). In Study 2, supervisors oversaw one to six therapists (M = 3.06, SD = 3.00), and their supervisees managed one to four cases (M = 2.44, SD = 1.03). In total, supervisors handled one to thirteen cases (M = 7.31, SD = 3.79) and participated in three to 39 supervision events (M = 21.31, SD = 11.25). Refer to Table 5 for further details.

The coding team comprised one postbaccalaureate student, eight doctoral students in clinical psychology (including the dissertation author), and two postdoctoral scholars in clinical psychology. Blind to study condition, the coders rated each supervision event using a subset of codes from the ACE-BOCS coding system (Chorpita et al., 2018). The coders evaluated code presence and extensiveness during moments in the supervision event, called “excerpts”, while also assessing the event holistically.

The initial phase of coder training involved a thorough review of coding procedures and codebook definitions, complemented by activities aimed at improving code recognition skills. Following this, coders independently coded a transcribed supervision event previously analyzed by the training team. Certification for coding proficiency was granted upon achieving 80% or higher agreement on excerpt-level codes and event-level extensiveness ratings for two consecutive supervision events, recognized as "gold standard" criterion events for training. After passing certification, coders participated in weekly meetings to review illustrative segments, compare codes and scores, rehearse coding aloud, assess discrepancies, and refine item content for construct consistency. Similar coding methods have demonstrated adequate reliability (Becker, Kim, Martinez et al., 2015).

Measures

Quality

Problem and practice classification involved four key decision-making activities within treatment engagement: considering problems, selecting problems, considering practices, and selecting practices. The "Considers Problem" and "Considers Practice" codes measured the depth of discussion and evidence

consideration, while the "Selects Problem" and "Selects Practice" codes assessed prioritization with supporting evidence.

These codes were assessed across five engagement domains for problems and eleven engagement strategies for practices. Within the five "REACH" domains (Relationship, Expectancy, Attendance, Clarity, and Homework), engagement problems encompass various facets of client-therapist interaction and therapy involvement (Becker & Chorpita, 2016). These domains offered a comprehensive framework for understanding and tackling engagement issues. Extensively documented in randomized controlled trials, practices addressing these problems ranged from versatile interventions effective across multiple domains to domain-specific strategies tailored to individual REACH domains (Becker et al., 2018). For instance, interventions enhancing therapeutic relationships may involve understanding identities, beliefs, and family dynamics, while those targeting expectancy may aim to elicit positive predictions about therapy outcomes. Similarly, tactics like regular appointment reminders and clear expectations can enhance attendance, whereas experiential learning activities facilitate practice outside of the session.

When evaluating activity quality, our initial step was to determine the presence or absence of the activity across the engagement targets. This was assessed through an event-level summary of its extensiveness, which is closely linked to performance quality. Extensiveness, defined as the intensity, depth, or “thoroughness” of a strategy (Hogue et al., 1996), is closely linked to performance quality (Garland, Hurlburt, et al., 2010). The use of extensiveness ratings to operationalize quality in therapy is a common approach within observational coding methods (Brookman-Frazee et al., 2021; Garland, Bickman, et al., 2010; McLeod & Weisz, 2010). Additionally, thorough consideration of evidence is widely acknowledged to enhance decision-making quality (Deming, 1989). Extensiveness ratings for each activity were coded observationally with the ACE-BOCS codebook (Chorpita et al., 2018), which has shown reliability and validity in measuring these competencies within engagement (Park et al., 2020). The ACE-BOCS codebook was informed by a knowledge management framework outlining essential steps for evidence-based decision-making (Graham et al., 2006). Scores range from 0 to 5, with an activity considered recognizable if it scored 2 or higher, indicating discussion that went beyond minimal

depth or detail. For instance, a "Considers Problem" rating of 1 suggests basic consideration, 3 indicates consideration with at least one source of evidence, and 5 reflects a thorough evaluation with multiple sources of evidence.

Next, these extensiveness ratings were used to determine a binary quality rating: 1 for adequately demonstrated and 0 for not demonstrated, ultimately indicating presence of the activity. "Considers Problem" and "Considers Practice" met quality standards if over two practices or problems scored 2 in extensiveness, implying the necessary consideration of multiple problems or practices, respectively. "Selects Problem" and "Selects Practice" met quality standards if only one or two problems or practices scored 2 or higher, reflecting a clear prioritization. Conversely, a score below quality was assigned if no selections were made or if three or more practices were chosen, indicating a lack of selectivity.

Max Extensiveness Score

For each of the four activities, the Max Extensiveness Score was calculated to represents the highest level of thoroughness achieved across all targets for each activity code (Considers Problem, Selects Problem, Considers Practice, Selects Practice) during each supervision event.

Effort

Effort was measured by the number of words spoken related to each assigned activity. This method counted the words spoken within each of the four activities during sessions, as identified by trained raters. An effort value of 0 was given if the activity was not observed in the event, whereas an effort value of 1 or greater indicated the activity was demonstrated. The total word counts for each instance the activity was observed in each excerpt were then aggregated across all instances within the session. Word count analysis is a standard method for assessing the duration of extended audio sequences in healthcare (Ziaei et al., 2016).

Total Event Words

This measure represents the total number of words spoken during the entire supervision session, from start to finish, irrespective of the subjects covered.

Total Event Minutes

This measure encompasses the entire duration of the supervision session, regardless of the topics discussed, spanning from the initiation to the conclusion of the supervision event.

Efficiency

Efficiency was measured using the ACE-BOCS code, "Efficiency," which assesses how promptly participants addressed engagement issues without unnecessary discussion. Efficiency was rated on a Likert scale ranging from 1 to 5, with higher scores indicating swift and organized decision-making. This score is not specific to any activity, problem, or practice but provides a holistic evaluation of event efficiency. Each supervision event received one Efficiency score.

Supervision Type

Supervision Type denotes the categorization of supervision events for each case, representing the chronological sequence of supervisory interactions from the initial event (1) to the fourth event (4) for a case. This classification enables us to capture the evolving nature of supervision events and their distinct characteristics as they correspond to the progression of each case through its various stages.

Supervisory Experience

Supervisory experience is defined by two temporal variables: the repetition of supervision activities and the weeks post-training. When categorized by supervision type, each repetition of supervision activity is referred to as a "Case." For example, First recorded supervision for Case 1, First recorded supervision for Case 2, and First recorded supervision for Case 3 are all within the first supervision type events. When all supervision types are included in the analysis, the focus shifts to the sequence of supervisory activities, where various cases and supervision types are intermixed. The repetition of supervision events (cases) indicates the sequential number of supervision sessions conducted by the supervisory dyad following the training event, reflecting repeated experience engaging in supervision activities. Weeks post-training measures the total elapsed time in weeks since the training event.

Data Analysis

Data Preparation

Part 1 of the study used data from both the experimental and comparison group. Part 2 used data exclusively from the experimental group. We considered mixed effects regression models to analyze our hierarchical dataset, which involved examining up to three nesting levels: cases nested within supervisees nested within supervisors. First, we checked how much each nesting level contributed to the variance in outcomes and whether it improved model fit. If a nesting level didn't add much to the variance or didn't make the model better, we simplified the model to a lower nesting level. Descriptive statistics and zero-order correlational analyses were calculated using SPSS Version 28. Otherwise, all other analyses were conducted in RStudio, Version 2023.12.1+402.

Part One

Part 1 established the foundation for identifying the metrics to be used in Part 2.

Q1: What is the reliability of quality and efficiency metrics?

The first aim of the study was to assess the reliability of activity quality metrics in evaluating the consideration and selection of problems and practices. Additionally, the study aims to evaluate the reliability of the Efficiency metric in assessing event-level efficiency. To examine interrater reliability, a stratified random sample of 30 (13.6%) supervision event transcripts were identified to be double-coded. To ensure that there existed enough transcripts to measure reliability for codes relevant to the competencies, 77% of double-coded transcripts were taken from the experimental group (n=65), who were specifically trained to use the study's competencies of interest. The remaining 23% (n=19) were selected from the comparison group. Two raters were chosen at random from a group of 11 possible raters to rate each supervision event transcript. The analysis was conducted on a sample of 84 transcripts of supervision events and 2 coders who independently rated each event.

We used Fleiss (Artstein & Poesio, 2008)' kappa (κ) (Fleiss, 1981) to evaluate interrater reliability for the four dichotomous quality metrics, since it is commonly applied to scenarios with

multiple raters or categories, when various raters assess different targets, and the raters are randomly selected from a larger pool, rather than being specifically assigned to each target. Fleiss' kappa accounts for chance agreement, giving a measure of agreement that truly reflects how much raters agree (Artstein & Poesio, 2008), and it quantifies the extent of agreement observed beyond what would be expected by chance alone. We used the default model, alpha (Cronbach), which assesses internal consistency by considering the average inter-item correlation. To calculate interrater reliability across all coders for the ordinal Efficiency metric, we used *ICC*s (model [2,1], consistency).

Fleiss' kappa and ICCs range from -1 to +1. Negative values indicate less agreement than expected by chance, with -1 suggesting no observed agreement and 0 indicating chance-level agreement. Positive values greater than 0 signify better-than-chance agreement, with +1 representing perfect agreement (Agresti, 2013). When interpreting Fleiss' kappa, values ≤ 0 signify no agreement, .01 to .20 represent poor agreement, .21 to .40 denote fair agreement, .41 to .60 indicate moderate agreement, .61 to .80 suggest good agreement, and .81 to 1.00 signify excellent agreement (Altman, 1990; Landis & Koch, 1977) According to established guidelines, ICCs less than .40 indicates poor agreement, .40 to .59 is considered fair, .60 to .74 is regarded as good, and .75 to 1.00 signifies excellent agreement (Cicchetti, 2001).

Based on pilot study of an earlier version of the codebook (Park et al., 2020), we predicted that the κ for the majority of the quality scores (i.e., 3 or more activity codes) would be within the acceptable range (i.e., $\kappa > 0.40$), per established cut-offs. We also predicted that the ICC for efficiency would be within the acceptable range (i.e., ICC >0.40), per established cut-offs (Koo & Li, 2016). Our null hypothesis was that the κs and ICC for most of the quality scores (3 or more) and for the efficiency score would not be 0.

Q2: What is the preliminary validity of the effort measure?

The second aim of the study was to determine preliminary construct validity of the effort measure. A linear mixed-effects regression model was fitted in R using restricted maximum likelihood

estimation (REML) with cases nested within supervisors and supervisors. Satterthwaite's method was utilized for t-tests. The model formula included Total Minutes as the dependent variable, with Total Words as predictor variables, and random intercepts specified for Supervisor, Supervisee, and CaseID. Next, we examined the association between Total event Words and Total Minutes. We predicted that these correlations would be significantly and positively correlated across each level of the sample at $\geq$0.45. Our null hypothesis was that across each level of the sample these correlations would equal 0.

Part Two

This dissertation investigated the reliability of quality, effort, and efficiency metrics, and then examined the associations between ordinal repetition of activities and passage of time with those quality and effort metrics. As such, it explored whether time or repetition is associated with improvement, deterioration, or no change in these metrics. We hypothesized that we would find a statistically significant positive slope (i.e., different than zero) in the experimental group for each metric, showing that *over time* (rather than all at once) supervisors in the experimental group improved in quality, effort, and efficiency within the specific competencies. We predicted that the Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type) would be significant, when compared with Weeks Post Training. Our null hypothesis was that the slope in the experimental group for each metric would not be different than zero, showing that over time supervisors in the experimental did not significantly improve in quality, effort, or efficiency within specific competencies. We initially included Supervision Type as a covariate in our analysis. Afterwards, we ran separate analyses for each Supervision Type, and examined the impact of Cases on these metrics.

Q3: Was quality associated with ordinal repetition of decision-making activities and/or the passage of time?

We examined if Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type) or Weeks Post Training predicted quality of the four activities: considering problems,

selecting problems, considering practices, and selecting practices. Given the potential for binary outcome variables to exhibit large effects (e.g., probabilities nearing 0% or 100%), we decided to use logistic regression models with a Firth penalty to assess the dichotomous quality outcomes in R. The Firth method is a penalized maximum likelihood estimation technique used in logistic regression to mitigate bias and stabilize parameter estimates, particularly in cases of rare events or highly imbalanced binary outcome data.

Q4: Was efficiency associated with ordinal repetition of decision-making activities and/or the passage of time?

For efficiency, we examined if Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type) or Weeks Post Training predicted session-level efficiency. A linear mixed-effects regression model was fitted using restricted maximum likelihood estimation (REML) with cases nested within supervisors and supervisors. Satterthwaite's method utilized for t-tests. The model formula included Efficiency as the dependent variable, with Repetition of Supervision Activities (or Cases, when categorized by Supervision Type) and Weeks Post Training as predictor variables, and random intercepts specified for Supervisor, Supervisee, and CaseID. Out of 341 supervision sessions, data for Efficiency were missing in two instances. To address this, the missing values were imputed using the grand mean score for efficiency in the CKS condition, which was determined to be 3.75.

Q5: Was effort associated with ordinal repetition of decision-making activities and/or the passage of time?

We examined if Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type) or Weeks Post Training predicted effort in the four competencies. Data analysis was conducted using a hurdle model, a statistical approach commonly employed for count data characterized by an excess of zeros. In Hurdle models, we are modeling excess zeroes separately from the rest of the data. They suggest that there is a hurdle (threshold) that separates the zero counts from the positive

counts. Different distributions are used to model these two components. The decision to use a hurdle model was driven by the nature of the data, which often contained numerous instances of zero counts due to non-occurrence of discussions within each competency. The hurdle model consists of two distinct components: a binary component estimating the likelihood of observing either zero or positive counts, and a count component modeling the frequency of positive counts. By employing the zero-inflated component of the hurdle model, we addressed the excess zeros present in the word count data. This component distinguished between true zeros (instances where no discussion occurred within the competency) and excess zeros (instances where discussion could have occurred within the activity but did not). The count component of the hurdle model, on the other hand, focused on modeling the frequency of positive counts, representing instances where discussion did occur. For our study, a truncated negative binomial distribution with a log link function was specified in the model's family argument, as it is well-suited for count data analysis, especially to address overdispersion. Separate hurdle models were fitted to the data, with the zero-inflated and count components estimated simultaneously. Data analysis was conducted using RStudio, Version 2023.12.1+402. We used the glmmTMB function from the glmmTMB package, with models fitted using maximum likelihood estimation. Random effects were assumed to be Gaussian on the scale of the linear predictor and integrated out using the Laplace approximation. Gradients were calculated using automatic differentiation. Within the mixed effects hurdle model, cases were nested within dyads.

Results

Part One

The descriptive statistics in Table 3a highlight several key aspects of supervision sessions form the study 1 sample. For the first supervision session, the average total minutes spent was 15.68, with a considerable range from 1.38 to 42.88 minutes, indicating quite varied session lengths in study 1. Weeks post-training had a median of 38.21 weeks, with some participants as recent as 15 weeks and others up to 92.14 weeks post-training.

Across the three supervision sessions, the total words dedicated to different activities varied. During the first supervision type, the mean number of words for "Considers Problem" was 920.17 (*SD* = 602.53). For "Considers Practice," the mean was lower at 217.77 (*SD* = 228.91). The "Selects Problem" activity had a mean of 93.03 words (*SD* = 158.32), while "Selects Practice" averaged 109.13 words (*SD* = 188.00). In the second supervision session, there was a notable decrease in words across all activities. "Considers Problem" dropped to a mean of 299.54 (*SD* = 482.58), "Considers Practice" decreased to 80.50 (*SD* = 106.90), "Selects Problem" further reduced to 20.75 (*SD* = 56.12), and "Selects Practice" averaged 76.54 words (*SD* = 118.01). The third supervision session showed a continued decline. The mean for "Considers Problem" was 185.31 (*SD* = 317.48), "Considers Practice" was 52.81 (SD = 119.21), "Selects Problem" dropped significantly to 1.85 (*SD* = 9.41), and "Selects Practice" was 38.81 (SD = 72.98). This trend indicates a consistent reduction in the number of words used across all categories as supervision sessions progressed.

Efficiency and extensiveness metrics showed similar trends. The efficiency score was 3.57 on average in the first supervision type, with a slight decrease in subsequent supervision types (3.32 and 3.42 for the second and third supervision types, respectively). Mean extensiveness scores for considering problems were relatively high in the first session (*M* = 3.90) and dropped in later sessions (mean was 2.00 in the second session and 1.19 in the third session). Selecting problems and practices showed similar declines. "Selects Problem" had a mean of 3.40 in the first supervision type, while "Considers Practice" and "Selects Practice" had mean scores of 2.70 and 2.90, respectively. In the second supervision session, these scores dropped, with "Considers Problem" at 2.00, "Selects Problem" at 1.11, "Considers Practice" at 1.39, and "Selects Practice" at 1.71. The third supervision session showed further decline, with "Considers Problem" at 1.19, "Selects Problem" at 0.73, "Considers Practice" at 0.69, and "Selects Practice" at 1.04.

The frequency of quality and effort in supervision events across the first, second, and third sessions in Table 4a shows distinct trends. In the first supervision session, 80.00% of events included "Considers Problem," but this dropped to 21.40% in the second and 15.40% in the third sessions. "Selects

Problem" was present in 63.30% of first sessions but fell to 21.40% in the second and was entirely absent in the third session. For practices, "Considers Practice" was present in 70.00% of first sessions, declining to 28.60% in the second and 11.50% in the third. "Selects Practice" was present in 50.00% of first sessions, 28.60% of second sessions, and 26.90% of third sessions.

Regarding effort, "Considers Problem" was present in 93.30% of first sessions, decreasing to 71.40% in the second and 57.70% in the third. "Selects Problem" showed a similar decline, from 60.00% in the first session to 28.60% in the second, and 7.70% in the third. "Considers Practice" was present in 76.70% of first sessions, 60.70% of second sessions, and 38.50% of third sessions. "Selects Practice" was present in 80.00% of first sessions, dropping to 57.10% in the second and 34.60% in the third.

Q1: What is the reliability of quality and efficiency metrics?

Fleiss' kappa was run to determine if there was agreement between trained raters' judgement of activity quality. There was excellent agreement between the raters' judgements on the problem activities, including Considers Problem ($\kappa = 0.89$, 95% CI [0.63, 1.09], $p < .001$) and Selects Problem ($\kappa = 0.83$, 95% CI [0.62, 1.04], $p < .001$). There was moderate agreement between the raters' judgements on the practice activities, including Considers Practice ($\kappa = 0.60$, 95% CI [0.39, 0.82], $p < .001$). Selects Practice ($\kappa = .451$, 95% CI [0.237, 0.67], $p < .001$) also showed moderate agreement, although its confidence interval contained 0. The interrater reliability of the *Efficiency* scores was assessed using the intraclass correlation coefficient (ICC) with a model of [2,1] and consistency type with both the untransformed data, presented. The ICC value was 0.77, indicating moderate interrater reliability. These results suggest that the coding system used by the coders on this measure was reliable, and that the ratings assigned by the coders were sufficiently consistent with each other and acceptable, per established cut-offs (Koo & Li, 2016). Please see Table 1 for additional details.

Q2: What is the preliminary validity of the effort measure?

Results from a mixed-effects linear regression model revealed a significant positive relationship between Total Words and Total Minutes in a supervision session. The predicted average minutes were

8.52 (95% CI [6.70, 10.34], $p = 0.219$), though this relationship was not statistically significant for the intercept. However, the predictor total words in a session showed a significant positive relationship with total minutes in the session, $\beta = 0.006$, $SE = 0.0001$, t (327.9) = 61.89, $p < 0.001$. Substantial variance in Total Minutes was attributed to clustering by case within supervisees within supervisors (ICC = 0.66), indicating that 66% of the total variance in Total Minutes was explained by differences between cases within supervisees and supervisors. The model demonstrated convergence at an REML criterion of 1576.1. Examination of scaled residuals revealed a range from -4.21 to 4.25, indicating no evident violation of the assumption of homoscedasticity. Please see Table 2 for additional details.

Additionally, the zero-order correlation analysis (one-tailed) conducted in SPSS revealed a strong positive relationship between the total number of words spoken during the event and the total duration of the event in minutes ($r = 0.95$, $p < .001$, $N = 337$). These findings indicate that as the number of words spoken increases, the event duration also tends to increase. Together these results suggest that word counts serve as a reliable indicator of the effort expended during the session.

Part Two

Descriptive statistics for continuous variables within the CKS condition across supervision events are summarized in Table 3b and categorical variables in Table 4b.

Total minutes per event across supervision types averaged 15.38 [95% CI: 14.38, 16.38], ranging from 2.80 to 47.80. The distribution of total minutes showed a moderate positive skew, with most values concentrated towards the lower end. While there were a few higher outliers, the overall shape had a lighter tail compared to a normal distribution.

For the first supervision session types, the mean weeks post-training was 57.81, with a range from 15 to 126.29 weeks and a standard deviation of 24.72, indicating substantial variability in the time since training. The median was 58.43 weeks. For the second supervision session, the mean weeks post-training increased to 61.65 weeks, with a range of 18.57 to 126.57 weeks and a standard deviation of 24.45, showing slightly less variability compared to the first session. The median was 61.00 weeks. In the third

supervision session, the mean weeks post-training further increased to 64.94 weeks, with a range from 22.71 to 127.43 weeks and a standard deviation of 24.26, like the previous session. The median was 64.14 weeks. This progression indicates an increase in weeks post-training across supervision sessions, with variability remaining consistent in the latter sessions.

Notably, in terms of verbal contributions in total supervision events, participants considering problems produced an average of 588.65 words (95% CI: 513.29, 664.02), while those selecting problems only produced an average of 45.41 words (95% CI: 35.63, 55.19). Similarly, participants considering practices averaged 213.05 words (95% CI: 175.99, 250.12), whereas those selecting practices averaged 87.45 words (95% CI: 72.41, 102.50). The word count distributions across all supervision events exhibited significant positive skewness, with most values concentrated towards the lower end. Additionally, a few exceptionally high outliers contributed to a long right tail in the distribution, resulting in a highly peaked and heavily tailed distribution compared to the normal distribution.

Efficiency scores across supervision events averaged 3.75 [95% CI: 3.64, 3.85] on a scale from 1 to 5. Most efficiency scores were clustered around the median (4.00), and the distribution was approximately symmetric with a slight negative skew and a relatively flat distribution compared to normal curve, with moderate variability and no extreme outliers.

In the context of supervision type, the mean total minutes per session during the First Supervision Events, which represent the initial recorded sessions for each case, was 19.81 [95% CI: 18.00, 46.28], notably higher than both Second Supervision Events (15.52) and Third Supervision Events (10.80). The distribution of total minutes in First Supervision Events displays a moderate positive skew, with most values concentrated towards the lower end, indicating relatively balanced session lengths with a few outliers. Conversely, Second Supervision Events, representing the subsequent supervision sessions, present a mean total minutes per session of 15.52 [95% CI: 13.83, 17.22], with a slightly more skewed distribution compared to the First Supervision Events. This skewness suggests greater variability in session durations, spanning a wider range. Third Supervision Events, marking the third set of supervision sessions, exhibit the shortest mean total minutes per session at 10.80 [95% CI: 9.54, 12.07]. The

distribution of total minutes in Third Supervision Events also shows a positive skew, albeit slightly more pronounced compared to the other types, indicating a tighter clustering of session lengths around the lower end.

Additionally, in terms of efficiency scores within supervision types, First Supervision Events demonstrate the highest mean efficiency (3.98) compared to both Second Supervision Events (3.67) and Third Supervision Events (3.60). The distribution of efficiency scores in First Supervision Events is approximately symmetric with a slight negative skew, with moderate variability and no extreme outliers. Similarly, Second Supervision Events display a similar distribution of efficiency scores but with a slightly lower mean, indicating slightly less efficiency on average compared to First Supervision Events. In contrast, Third Supervision Events also exhibit a negative skew in efficiency scores but with a slightly less pronounced skewness compared to the other types. The shape of the distribution suggests tighter clustering of efficiency scores around the median, with fewer extreme values.

As indicated in Table 3b, for the first supervision session, "Considers Problem" had a mean extensiveness score of 4.56, and "Selects Problem" had a mean of 4.26, indicating high extensiveness. "Considers Practice" and "Selects Practice" had lower mean scores of 3.14 and 3.11, respectively. In the second supervision session, these scores dropped: "Considers Problem" at 2.06, "Selects Problem" at 1.24, "Considers Practice" at 1.59, and "Selects Practice" at 1.96. The third supervision session showed further declines, with "Considers Problem" at 1.14, "Selects Problem" at 0.56, "Considers Practice" at 0.50, and "Selects Practice" at 0.55. Overall, the "Considers Problem" and "Selects Problem" distributions were highly skewed to the left with many high scores, whereas "Considers Practice" and "Selects Practice" distributions are less skewed, more spread out, and closer to normal distributions in the first supervision type.

Table 4b illustrates the frequency and percentage distribution of quality and effort within supervision events across the CKS condition, segmented by activity and supervision type. The frequencies of quality presented within the First, Second, and Third Supervision sessions depict no meaningful change. The frequencies of quality presented across the First, Second, and Third Supervision

sessions depict a noticeable pattern change over time. In the First Supervision, most cases showed quality Considers Problem (96.60%) and Considers Practice (88.10%), while fewer exhibited qualities like Selects Problem (78.80%) and Selects Practice (73.70%). As the sessions progressed to the Second and Third Supervision type, there was a notable shift. The percentage of cases demonstrating quality decreased significantly across the board, with a more pronounced decrease in qualities related to problem selection compared to those linked to problem consideration and practice. For instance, while the percentage of cases demonstrating quality within Considers Problem remained relatively stable between the First and Third Supervision (96.60% to 86.00%), the percentage for quality within Selects Problem decreased substantially from 78.80% to 14.00%. This suggests a potential trend towards a decrease in the engagement of certain activities over time, particularly those involving problem selection.

Q3: Was quality associated with ordinal repetition of decision-making activities and/or the passage of time?

To prepare the data, initially, we assessed the linearity assumption for all quality outcome variables. This assumption suggests that as the continuous independent variable, Weeks Post Training, increases by one unit, the log odds (logit) of the dependent variable should consistently change by a constant amount. For *Considers Problem, Selects Problem*, and *Considers Practice*, no significant linear associations were observed across the tested models ($p > 0.05$ for all). However, for *Selects Practice*, the relationship between the natural log transformation of *Weeks Post Training* by *Weeks Post Training* and *Selects Practice* was significant ($\beta = -0.046$, Wald = 7.115, df = 1, $p = 0.008$), indicating a non-linear association. Subsequent power transformations for *Weeks Post Training* were considered to address nonlinearity, with a power transformation of approximately 2 suggested based on the findings. Analysis for *Selects Practice* was run with and without the power-transformed Weeks Post Training, for comparison. No outliers were detected for *Considers Problem, Selects Problem*, and *Considers Practice, and Selects Practice*, as indicated by their absence on casewise plots. Omnibus tests of model coefficients

revealed statistically significant chi-square values (Chi-square = 21.144, df = 2, $p < .001$) for all models, suggesting adequate model fit.Q2

Table 6 shows the results of all the quality outcomes, with Supervision Type, Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type), and Weeks Post Training included as covariates. Logistic regression models with Firth corrections revealed consistent significance for Supervision Type across all quality outcomes, with odds ratios (ORs) ranging from 0.10 to 0.29. These results indicate that as Supervision Type increases, the likelihood of quality presence in these competencies significantly decreases by 71% to 90.0%**.** In other words, the odds will decrease by a factor of 0.10 to 0.29 for each one-unit increase in Supervision Type. I obtained the percentage by subtracting one from the odds ratio and multiplying by 100. In contrast, Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type) did not show a significant association with quality across all metrics. Similarly, except for the "Selects Practice" activity, Weeks Post Training did not show a significant association with quality. The odds ratio (0.98) for "Selects Practice" indicated a marginal decrease in the likelihood of quality presence with each unit increase in Weeks Post Training.

The results presented in Table 7 demonstrate the association of Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type)and Weeks Post Training on the likelihood of activity quality in first, second, and third supervision types. Across all three supervisions types, Repetition of Supervision Activities did not exhibit statistically significant associations with activity quality, suggesting that variations in Repetition of Supervision Activities were not significantly related to changes in activity quality across the different stages of supervision. Apart from the "Selects Practice" activity in the second supervision type, which showed a negligible odds ratio (0.99), Weeks Post Training did not demonstrate a significant association with quality. When categorized by Supervision Type, Cases were not significantly related to changes in activity quality or effort with repetition of activities.

Figures 1a through 1c illustrates the zero-order proportions of quality in problem consideration and selection, as well as practice consideration and selection, within each Supervision Event Type. For first supervision event types, most supervision events demonstrated a high percentage of quality in

decision-making activities (over 65%, many reaching 100%). In contrast, for the second supervision event types, quality was observed less frequently (mostly in the 30% to 50% range), and even less so for the third supervision event type (with most at 0%, and a few ranging from 10% to 25%).

Q4: Was efficiency associated with ordinal repetition of decision-making activities and/or the passage of time?

Shown in Table 9, results from a mixed-effects linear regression model revealed that immediately after training, predicted average efficiency was 3.88, 95% CI [3.39, 4.37], $p < 0.01$. Substantial variance in Efficiency was attributed to clustering by case within supervisees within supervisors (ICC = 0.45), indicating that 45% of the total variance in Efficiency was explained by differences between cases within supervisees and supervisors. Supervision Type significantly predicted Efficiency, indicating a decrease as supervision events progressed by 0.19 ($b_1 = -0.19$, t (311.70) = -3.41, $p < 0.01$). Weeks Post-Training and Repetition of Supervision Activities were not significant predictors of Efficiency. The model demonstrated convergence at a REML criterion of 920.3. Examination of scaled residuals revealed a range from -2.68 to 1.92, indicating reasonable overall fit of the model.

Table 9 presents the results of mixed-effects linear regression models predicting efficiency across the first, second, and third supervisions separately. Neither Cases nor Weeks Post Training show significant associations with efficiency in any of the supervision types. Figure 3 illustrates the zero-order proportions of efficiency within each Supervision Event Type.

Q5: Was effort associated with ordinal repetition of supervision activities and/or the passage of time?

Table 10 shows the results of all the effort outcomes from the hurdle analysis with Session Event Type, Repetition of Supervision Activities, and Weeks Post Training as covariates. The data are divided into count models and zero-inflated models. The Incidence Rate Ratios (IRRs) are used to describe the relationship between a predictor variable and the rate of occurrence of an event over time or another unit of measurement. An IRR of 1 indicates no association between the predictor and the outcome rate, an IRR

greater than 1 indicates an increased rate of the outcome with a one-unit increase in the predictor, and an IRR less than 1 indicates a decreased rate of the outcome with a one-unit increase in the predictor. In the context of the hurdle model, the OR can be used to interpret the likelihood of an excess zero count (i.e., the odds that the outcome is zero as opposed to a positive count). Odds Ratios (ORs) less than 1 indicate a decrease in the odds of zero counts, while ORs greater than 1 indicate an increase. Both OR and IRR reflect multiplicative changes and are calculated by exponentiating the regression coefficient (β), in which e^{β} . The significance of each predictor was determined by *p*-values. This analysis helps understand how different predictors affect the presence of counts and the likelihood of zero counts in the context of supervision type, repetition of supervision activities, and weeks post training.

In the count model, for supervision type, the IRR was 0.50 for "Considers Problem," indicating the effort count is halved compared to the reference group for a unit increase in supervision type (β = -0.69, SE = 0.06, $p < .001$). The IRRs for "Selects Problem," "Considers Practice," and "Selects Practice" were 0.94, 0.87, and 1.08, respectively, were all non-significant, p = .129 to .658. Repetition of supervision activities showed slight, non-significant decreases in expected effort count across the four activities, with IRRs ranging from 0.91 to 0.98. Weeks post-training had IRRs of 1.00 to 1.01, indicating no significant change in expected effort count. The count model also revealed that among participants displaying effort initially post-training, those considering problems were expected to speak an average of 1,924 words (IRR = 1924.41, 95% CI [1365.04, 2712.99], β = 7.56, SE = 0.18, $p < .001$), whereas those selecting problems were expected to voice about 120 words (IRR = 119.72, 95% CI [67.72, 211.65], β = 4.79, SE = 0.29, $p < .001$). Similarly, individuals considering practices averaged 453 words (IRR = 453.27, 95% CI [269.80, 761.51], β = 6.12, SE = 0.26, $p < .001$), and those selecting practices expressed an average of 110 words (IRR = 109.88, 95% CI [67.21, 179.64], β = 4.70, SE = 0.25, $p < .001$), all statistically significant. Applying an average speaking rate of 140 wpm, supervisory dyads who demonstrated an activity typically spent about 13.74 minutes considering problems, 0.86 minutes selecting problems, 3.24 minutes considering practices, and 0.79 minutes selecting practices.

In the zero-inflated model, the odds ratios (ORs) for supervision type ranged from 4.88 to 10.74, indicating an increased likelihood of zero counts with each unit increase in supervision type across all supervision activities. Specifically, for "Considers Problem," an OR of 4.88 (β = 1.56, SE = 0.27, p < .001) means that each unit increase in supervision type increases the likelihood of observing zero counts by a factor of 4.88. For example, a supervisory dyad under the second supervision type is expected to have no instances of "Considers Problem" 4.88 times more often compared to the first supervision type. For "Selects Problem," an OR of 10.74 (β = 2.37, SE = 0.28, p < .001) indicates that each unit increase in supervision type increases the likelihood of zero counts by a factor of 10.74. For "Considers Practice," an OR of 6.07 (β = 1.80, SE = 0.25, p < .001) suggests that each unit increase in supervision type increases the likelihood of zero counts by a factor of 6.07. For "Selects Practice," an OR of 6.35 (β = 1.85, SE = 0.25, p < .001) implies that each unit increase in supervision type increases the likelihood of zero counts by a factor of 6.35. Repetition of supervision activities had ORs of 1.06 to 1.14, indicating a slight, non-significant increase in the odds of zero counts across all four activities, p = .298 to .612. Weeks post-training had ORs close to 1 (0.99 to 1.01), suggesting minimal impact on the odds of zero counts that were all not significant, p = .208 to .918.

Tables 11a through 11c reveal the likelihood of activity effort based on the repetition of supervision activities and weeks post-training for the first, second, and third supervision types.

For "Considers Problem", the count model for first supervision type showed a significant baseline count (IRR = 1395.36, 95% CI [931.77, 2089.61], β = 7.24, SE = 0.20, z = 35.14, p < .001). A unit increase in the predictor "Case" resulted in a 7% increase in the expected count of effort (IRR = 1.07, 95% CI [1.02, 1.12], β = 0.06, SE = 0.02, p = .005). Furthermore, each additional week post-training marginally decreased the expected count of effort by 1% (IRR = 0.99, 95% CI [0.99, 1.00], β = -0.01, SE = 0.00, p = .002). The zero-inflated model results were not interpretable due to extremely wide confidence intervals and non-significant p-values.

For "Selects Problem," the effect of Case (IRR = 0.95, 95% CI [0.87, 1.04], β = -0.05, SE = 0.05, p = .242) and weeks post-training (IRR = 1.00, 95% CI [0.99, 1.01], β = 0.00, SE = 0.00, p = .551) on the

expected count of effort for first supervision type were not significant. The zero-inflated model indicated extremely low odds of zero counts (OR = 0.00, β = -12.60, *SE* = 3.56, $p < .001$), with non-significant effects for both predictors in the first supervision type.

For the "Considers Practice" condition, the count model demonstrated a significant baseline count (IRR = 307.5, 95% CI [187.93, 503.14], β = 5.73, *SE* = 0.30, $p < .001$). Neither the effect of "Case" (IRR = 0.96, 95% CI [0.88, 1.04], β = -0.04, *SE* = 0.04, $p = .319$) nor weeks post-training (IRR = 1.00, 95% CI [0.99, 1.01], β = 0.01, *SE* = 0.01, $p = .434$) were significant. The zero-inflated model showed significantly low odds of zero counts (OR = 0.00, β = -13.87, *SE* = 5.83, $p = .017$) at baseline, with non-significant effects for both predictors, case, and weeks post training.

For the "Selects Practice" condition, the count model revealed a significant baseline count (IRR = 156.18, 95% CI [91.45, 266.71], β = 5.05, *SE* = 0.27, $p < .001$). A unit increase in "Case" slightly decreased the expected count of effort, but this effect was marginal (IRR = 0.91, 95% CI [0.82, 1.00], β = -0.10, *SE* = 0.05, $p = .049$). Weeks post-training had no significant effect on the expected count of effort (IRR = 1.00, 95% CI [0.99, 1.01], β = 0.00, *SE* = 0.01, $p = .891$). The zero-inflated model showed significantly low odds of zero counts (OR = 0.00, β = -11.04, *SE* = 3.71, $p = .003$) at baseline, with non-significant effects for both predictors, case, and weeks post training.

Similarly, the results for the second supervision type revealed significant baseline counts across all activities: "Considers Problem" (IRR = 225.66, 95% CI [131.81, 386.35], $p < .001$), "Selects Problem" (IRR = 98.6, 95% CI [23.67, 410.77], $p < .001$), "Considers Practice" (IRR = 433.55, 95% CI [210.91, 891.24], $p < .001$), and "Selects Practice" (IRR = 95.42, 95% CI [44.52, 204.52], $p < .001$). However, the effects of "Case" and weeks post-training were not significant in any activity, indicating that these predictors did not significantly impact the expected count of effort during the second supervision type. In the zero-inflated model, the odds of zero counts were not significantly affected by either "Case" or weeks post-training for any condition. For "Considers Problem," "Case" (OR = 1.06, $p = .583$) and weeks post-training (OR = 0.99, $p = .692$) were non-significant. For "Selects Problem," "Case" (OR = 0.97, $p = .852$) and weeks post-training (OR = 1.02, $p = .423$) were non-significant. For "Considers Practice," "Case"

(OR = 1.01, p = .889) and weeks post-training (OR = 1.00, p = .691) were non-significant. For "Selects Practice," the intercept indicated low odds of zero counts (OR = 0.15, p = .036), but "Case" (OR = 0.98, p = .894) and weeks post-training (OR = 1.01, p = .449) were non-significant.

For the third supervision type, the count model showed significant baseline counts across all conditions: "Considers Problem" (IRR = 211.04, 95% CI [111.26, 400.32], p < .001), "Selects Problem" (IRR = 29.09, 95% CI [10.44, 81.04], p < .001), "Considers Practice" (IRR = 380.75, 95% CI [210.63, 688.27], p < .001), and "Selects Practice" (IRR = 136.61, 95% CI [63.33, 294.67], p < .001). The effects of "Case" and weeks post-training were generally not significant, except for a slight increase in expected counts for "Considers Practice" (Case IRR = 1.07, p = .006) and a decrease for weeks post-training (IRR = 0.98, p = .001). In the zero-inflated model, the odds of zero counts were not significantly affected by either "Case" or weeks post-training for any condition, indicating that these predictors did not significantly influence the likelihood of zero counts within the third supervision type.

Figures 3a through 3c shows zero-order box plots for all the effort outcomes across dyad sessions within supervision type.

Discussion

Part One of this dissertation explored the reliability of quality, effort, and efficiency metrics within the framework of a decision-support system integrated into clinical supervision. Part Two investigated how the repetition of activities and the passage of time relate to these quality and effort metrics, aiming to determine whether improvements, deteriorations, or no changes in these metrics were associated with time or repetition.

Part One

In Part One, we predicted that Fleiss' κ for most of the quality scores and ICC for efficiency will be within the acceptable range (i.e., κ >0.40 and ICC >0.40), per established cut-offs (Altman, 1990; Cicchetti, 2001; Koo & Li, 2016; Landis & Koch, 1977). Reliability analysis revealed excellent

agreement for problem competencies (Fleiss' κ = .887-.829), moderate agreement for practice competencies (Fleiss' κ = .601-.451), and very good interrater reliability for overall event efficiency (ICC = 0.765). These findings suggest that the coding system used by the raters was reliable and the ratings were consistent with established criteria (Koo & Li, 2016) and showed greater reliability overall compared to prior study examining reliability of the ACE-BOCS coding system using extensiveness scores rather than summary dichotomous ratings (Park et al., 2020). This was expected, as consolidating multiple codes typically improves reliability (Heyman et al., 2021).

Implications

The difference in reliability between problem competencies and practice competencies, as indicated by Fleiss' κ coefficients, suggests varying levels of consistency in the coding process for these two types of competencies. The excellent agreement for problem competencies indicated a high level of consensus among raters when assessing problem-related behaviors. In contrast, the moderate agreement for practice competencies suggests a somewhat lower level of agreement among raters when evaluating practice-related behaviors.

The "considers" activity code demonstrated better reliability when evaluating a target problem than when assessing another target practice. Several factors may contribute to this discrepancy in reliability in supervision events overall. For example, Table 3a illustrates that supervisory dyads typically considered more problems than practices. For "Considers Problem," the mean number of problems considered was 1.88 [95% CI: 1.74, 2.01], whereas "Considers Practice" had a mean of 1.25 [CI: 1.10, 1.40]. For "Selects Problem," the mean was 0.66 [CI: 0.58, 0.74], compared to "Selects Practice," with a mean of 0.59 [95% CI: 0.52, 0.66]. Both "Selects Problem" and "Selects Practice" had zero standard deviation, indicating consistent performance. Practice-related activities involved fewer words and decisions, providing less material for coders to evaluate accurately. This difference in material could affect reliability estimates. Analyzing the number of targets or excerpts in high versus low-quality groups for the four activities could reveal further notable differences, warranting additional research.

We also tested the hypothesis that measures of total word counts and total event duration these correlations will be significantly and positively correlated. Mixed-effects linear regression and zero-order correlational analysis showed significant positive relationships between word counts and event duration ($r = 0.95$, $p < .001$). . This finding suggests that longer events tend to involve more words spoken, reflecting a greater degree of engagement and active discussion regarding the competencies under consideration. The correlation aligns with previous research in healthcare that has used word counts as proxies of effort (Ziaei et al., 2016)., further corroborating the validity of this approach in assessing activity effort within clinical supervisory contexts. This indicates that word counts can serve as a useful metric for quantifying effort.

Limitations

The descriptive statistics in Table 4a show the first, second, and third supervision sessions in study 1 indicating that it will be more challenging to be reliable in the first and third supervision types, given the high and low frequency of these behaviors, respectively. For "Considers Problem," the behavior is present in 80.00% of first sessions but drops to 15.40% in the third stage. Similarly, "Selects Problem" is present in 63.30% of first sessions but declines to 0.00% in later stages sessions. For practices, "Considers Practice" is present in 70.00% of first sessions but decreases to 11.50% by third supervision, while "Selects Practice" is present in 50.00% of first sessions and falls to 26.90% by third supervision. Given our primary interest in the first supervision stage, where the behaviors of interest are most expected and desired to occur, it is possible that the reliability of these types of transcripts could be much lower in that stage.

Furthermore, given that our primary interest lies in applying these codes to the first supervision stage of *Study 2*, where the behaviors of interest are most expected to occur, rather than Study 1, it is possible that the reliability of these types of transcripts could be much lower due to the restricted range in the study sample of study 2. In Study 2, for the first supervision session, "Considers Problem" was present in 96.60% of cases, and "Selects Problem" was present in 78.80% of cases. Similarly, "Considers

Practice" was present in 88.10% of cases, and "Selects Practice" was present in 73.70% of cases. The effort metrics also show high presence rates, with "Considers Problem" present in 98.30% and "Selects Practice" present in 84.70% of cases. In contrast, the first supervision session in Study 1 shows more variability: "Considers Problem" was present in 80.00% of cases, and "Selects Problem" was present in 63.30% of cases. "Considers Practice" was present in 70.00% of cases, and "Selects Practice" was present in 50.00% of cases. The effort metrics also indicate more variability, with "Considers Problem" present in 93.30% and "Selects Practice" present in 80.00% of cases. The reduced variability in Study 2, suggests that the reliability of these codes will be lower compared to Study 1. The high frequency of presence and limited instances of absence in the behaviors of interest can lead to difficulties in distinguishing true quality differences, ultimately impacting the reliability of the assessments when applied to Study 2.

Our study found a significant positive correlation between total word counts and event duration; however, there are several limitations to consider. Firstly, the correlational nature of the analysis precludes establishing causality, and it's plausible that other factors not accounted for in our study may influence both event duration and word counts. Additionally, the use of word counts as a proxy for effort may not capture the qualitative aspects of engagement or the depth of discussion. Furthermore, our study focused solely on word counts without considering other potential indicators of effort or engagement, such as non-verbal communication cues or participant perceptions. Therefore, whereas word counts can provide valuable insights into the level of activity during supervision sessions, they should be interpreted cautiously and in conjunction with other measures to ensure a comprehensive understanding of activity effort within clinical supervisory contexts.

It's important to note that quality, effort, and efficiency extends beyond thorough consideration or selection of problems or practices; it also encompasses the appropriateness of the selected problem and practice for the specific case. However, this distinction was not addressed in our study, as we defined any thorough consideration or selection of problems or practices as high quality, irrespective of their relevance to the case at hand. This approach may overlook potential variations in the quality of problem or practice selection based on their appropriateness for individual cases.

The prevalence of many zero ratings for quality predictors in the combined sample of supervision types collapsed together led to adoption of binary ratings indicating the presence or absence of quality. In addition to the threshold issue where most supervision events exhibited presence of quality and effort, the use of a dichotomous quality variable further restricts our understanding of the extent and depth of quality within supervisory discussions. Consequently, we could only report percentages in Part 2, which may not fully capture the intricacies of quality variations across sessions. Furthermore, this dichotomous variable precluded the creation of a composite efficiency score, limiting our ability to comprehensively assess supervisory quality. Consequently, we were unable to directly examine the ratio of quality and effort for each supervisory dyad per session, which could have provided valuable insights into efficiency for part 2 of the study.

Future Directions

Several avenues for future research present themselves based on the findings and limitations of our study. Firstly, given the discrepancy in reliability between problem and practice competencies, further investigation into the underlying factors contributing to this difference is warranted. Understanding why raters exhibit higher consensus when assessing problem-related behaviors compared to practice-related behaviors could provide valuable insights into the nature of these competencies and the coding process. Additionally, exploring alternative coding schemes or approaches to enhance the reliability of practice activity assessment may be beneficial.

Furthermore, whereas our study demonstrated a significant positive correlation between word counts and event duration, future research should aim to validate the use of word counts as a measure of activity effort in clinical supervisory contexts. This could involve examining the relationship between word counts and other indicators of engagement or performance quality to ensure a comprehensive understanding of activity effort. Additionally, exploring the role of silence in supervisory interactions and its impact on event outcomes by using event duration and time stamps could provide valuable insights. Conducting qualitative analyses to examine the content and function of silent periods during supervision

would be essential, exploring how moments of reflection or contemplation contribute to problem-solving and decision-making processes.

Moreover, addressing the limitations identified in our study is crucial for advancing research in this area. Future studies should explore alternative methods for assessing quality in supervisory discussions, considering the appropriateness of problem and practice selection for individual cases. Additionally, efforts to develop a composite efficiency score that combines quality and effort measures would provide a more holistic understanding of supervisory effectiveness.

Overall, future research should strive to refine and expand upon the methodologies employed in our study to better capture the complexities of clinical supervision and enhance the validity and reliability of activity assessment. By addressing these areas, future studies can contribute to the ongoing improvement of clinical supervisory practices and ultimately enhance the quality of mental health services delivery.

Part Two

In Part Two of the study, we aimed to determine whether improvements, deteriorations, or no changes in these metrics were linked to time or repetition. We hypothesized that the experimental group would exhibit a statistically significant positive trend for each metric, indicating that supervisors in this group improved their performance in quality, effort, and efficiency with practice in specific competencies. Consequently, we expected the Repetition of Supervision Activities (i.e., Cases, when categorized by Supervision Type) to be significant, but not Weeks Post Training, suggesting that changes in effort, quality, and efficiency were driven by repetition of supervisory activities rather than the passage of time. However, our analyses of effort, quality, and efficiency did not support this hypothesis.

We observed declines in efficiency, quality presence, and effort likelihood with each successive one-unit increase in Supervision Type, indicating notable differences in performance between supervisory stages. With each unit increase in Supervision Type, there was a notable decrease in the likelihood of quality in decision-making activities (ranging from 71% to 90.0%) and effort (ranging from 3% to 66%).

This decrease was particularly significant for quality and effort in considers problem (90% and 66%, respectively), as well as effort in considers practice (35%). The predicted efficiency also decreased by 0.19 units for every one-unit increase in Supervision Type. Given that in different stages of supervision, certain activities may be more typical and appropriate for a particular stage, likely contribute to the observed differences. For example, problem finding would ideally be more prevalent in the first supervision stage compared to subsequent supervision stages, which ideally focus on different decision-making activities like reviewing performance in practice implementation — aspects not explored in this study.

More importantly, cases did not significantly predict decision-making activity outcomes within each supervision type, suggesting limited association with Repetition of Supervision Activities after training. Weeks Post Training generally had negligible effects, also indicating no meaningful impact on performance outcomes over time. Overall, these results highlight sustained but not improved performance levels within supervision stages without additional intervention.

Despite significant variability in weeks post-training, with values ranging from 15.00 to 126.29 weeks in the first supervision session and 22.71 to 127.43 weeks in the third, the presence of quality and effort in supervision events remained stagnant. This stagnation occurred even though the time since training varied widely among participants, suggesting that the length of time post-training did not significantly influence the quality and effort observed during supervision. Specifically, quality metrics like "Considers Problem" and "Selects Problem" showed high initial presence but declined in later sessions, regardless of the weeks post-training. Similarly, effort metrics followed the same pattern, indicating that factors other than time since training might play a more critical role in sustaining quality and effort in supervisory practices.

Implications

The finding that more activities were observed in supervision stage one than in stages two and three, where they were less expected, highlights a positive alignment of supervisory practices with the

natural progression of case management. This suggests that supervisors are effectively focusing on critical activities such as considering problems and selecting practices at the beginning of supervision. By concentrating on problem identification and initial planning during the early stages, supervisors ensure that foundational issues are addressed upfront. This allows for a more efficient and effective supervisory process, as the groundwork established in stage one supports smoother transitions into subsequent stages. Once these crucial initial tasks are completed, supervisors can then move on to other important activities, such as implementation review and performance assessment, which are more appropriate for later stages. This structured approach not only enhances decision-making and quality outcomes but also ensures that each stage of supervision is aligned with the evolving needs of the case, leading to better overall results.

At first glance, the stagnation in the percentage of dyads demonstrating quality, effort, and efficiency in key decision-making activities within supervision stage one may suggest the need for targeted interventions or training programs to address these challenges and ensure consistent quality in supervisory practices across all stages. The prevalence of many zero ratings for quality predictors in the combined sample of supervision types led to the adoption of binary ratings indicating the presence or absence of quality. This approach, however, limits the nuance of quality assessment, particularly in the first supervision stage, which showed nearly 100% presence of quality and effort. Thus, there appears to be a threshold issue where most supervision events showed the expected behaviors, making it difficult to distinguish true quality. This raises a critical concern about the validity of our quality measure. The current metric merely indicates the presence of an activity rather than genuinely assessing its quality, as the threshold was set too low.

The finding that quality and effort in supervision events remained stagnant despite significant variability in weeks post-training suggests that simply allowing more time to pass after training does not necessarily enhance or sustain the quality and effort in supervisory practices. This challenges the assumption that more experience or time since training inherently leads to better supervision outcomes. Incorporating booster events via expert consultation in decision-making supports could enhance retention or increase learning in skill building in activities, as evidenced by prior research examining consultation

(Frank et al., 2020; Lyon et al., 2015; Ngo et al., 2011; Petry et al., 2012). Furthermore, supervisory dyad's significant emphasis on classifying and selecting problems compared to classifying and selecting practices raises questions about the underlying reasons for this disparity. Overall, these findings provide valuable insights into the complexities of supervisory practices and point towards the need for ongoing research and intervention efforts to optimize the quality and efficiency of clinical supervision in mental health settings.

Limitations

In addition to the limitations discussed regarding the measures used in Part 1, there are further considerations in this study. The declining frequency of dyadic supervision events beyond the initial nine events pose a significant limitation in our analysis. As the study progresses, the diminishing number of data points from events 10 to 14 presents a challenge in accurately assessing temporal changes. This limitation arises from the reduced availability of data, making it challenging to capture the subtleties of evolving supervisory interactions. Consequently, the study's capacity to effectively analyze and interpret longitudinal changes may be compromised, emphasizing the importance of data sufficiency in examining temporal dynamics in research. Additionally, the analysis did not incorporate confounding variables such as burnout or caseload, which could potentially influence supervisory dynamics. Failure to consider these factors may introduce bias and limit the comprehensiveness of our findings. Future research should aim to address these confounding variables to provide a more nuanced understanding of the factors affecting supervisory quality and effort.

Future Research

Building upon the findings of this study, future research should focus on more accurately assessing quality, effort, and efficiency in decision-making activities within the first supervision stage. Given that Figure 1a showed nearly all activities rated at 100% quality, it is crucial to raise the quality bar by revisiting the measure of extensiveness or implementing a higher threshold for quality. By setting more

stringent criteria for quality and effort, we can better distinguish true quality and ensure that our assessments reflect meaningful variations in supervisory practices.

To address this, we should establish a higher standard for quality in S1, using a more nuanced metric that evaluates the extensiveness and effectiveness of the observed activities. This approach would provide a more accurate reflection of what constitutes quality in supervision events. Inspection of the descriptive statistics of the extensiveness scores before transformation into the dichotomous quality variable revealed a significant range and variability in engagement, suggesting that an extensiveness threshold above the sample average might be more nuanced and make more sense. As indicated in Table 3b, for the first supervision session, "Considers Problem" had a mean extensiveness score of 4.56, and "Selects Problem" had a mean of 4.26, indicating high engagement. In contrast, "Considers Practice" and "Selects Practice" had lower mean scores of 3.14 and 3.11, respectively. These findings imply that setting an extensiveness threshold around these higher averages, such as 4.00 or higher, could better capture the depth and quality of engagement in supervisory activities. Alternatively, future research may consider using the max extensiveness scores without transformation or dichotomization. Using the extensiveness scores without transformation is feasible, particularly for "Considers Practice" and "Selects Practice," given their closer-to-normal distributions. For the more skewed "Considers Problem" and "Selects Problem," careful application of robust statistical methods and thorough reporting of descriptive statistics will help ensure that the analysis accurately reflects the underlying data and provides meaningful insights into supervisory quality. By adopting more detailed variable metrics, we may better distinguish between varying levels of supervision quality and provide more targeted feedback for improvement.

Although it is worth noting that even a highly extensive supervision activity does not necessarily equate to a high-quality one. To truly understand quality, we could consider the ultimate effects these supervisory activities are intended to produce. For example, if a newly extensively defined quality for the first supervision stage (S1) did not correlate with performance in the first therapy stage (T1), this would suggest that our metric might not be capturing quality accurately. Instead, it may simply be indicating whether an activity occurred, not how well it was performed.

To determine true quality, we might consider the ultimate goals of these activities, such as accurately identifying and addressing the problems that the youth face. For instance, if supervisors who extensively considered the problem were more likely to identify the youth's actual issue, as secretly reported by the youth themselves, this would demonstrate true quality. An experiment where raters did not see the graphical report of the REACH domains of concern but accurately matched what the youth reported as their problem would support this. This alignment would suggest that thorough problem consideration leads to accurate problem identification, a key aspect of quality supervision.

To further explore ways to define and measure quality, we need reliable indicators that go beyond the mere presence of behaviors. As another example, supervisory activities could be assessed not just by whether it occurred, but by the appropriateness and effectiveness of the selected practice, as indicated in the treatment events that immediately follow. Ultimately, to feel confident that our measures truly reflect quality, we must ensure they lead to better outcomes, such as accurately identifying and addressing the youth's issues and improving overall therapeutic performance.

Longitudinal studies that track supervisory dyads' performance over extended periods, with increased repetition of supervision activities (i.e., cases, when categorized by supervision type), could provide valuable insights. These studies may reveal an initial plateau in performance, followed by significant improvement in learning and skill development beyond the scope of our current evaluation range. This extended observation period would allow us to capture long-term trends and better understand how repeated practice activities contribute to sustained improvements in supervision quality and effectiveness.

Additionally, investigating the impact of organizational factors, such as workload and institutional support (Msuya & Kumar, 2022), on supervisory efficiency could enrich our comprehension of this multifaceted phenomenon. We could do this by including in the analysis the ratio of off-topic discussions and on-topic discussions in supervision, interruptions, and other efficiency-challenging behaviors between supervisors and supervisees within supervision sessions. Similarly, considering the subjective perception of effort or motivation experienced by supervisory dyads may offer further insights.

Despite the potential benefits of decision support systems, mastering their usage and achieving proficiency typically requires considerable time and effort. Given the vast amount of knowledge to be acquired within limited timeframes and retained over the long term, paced learning emerges as a promising strategy to address the decrease in quality and effort observed across supervision events (Chorpita et al., 2021). Paced learning within mental health service systems entails identifying the necessary workforce competencies and devising tailored learning strategies to enable individuals to acquire them at an appropriate pace. Paced learning ensures a continuous cycle of learning, introducing new training components as individuals master preceding ones.

Another avenue for future research involves comparing computer-generated efficiency metrics with those derived from human assessment. Exploring methodologies to assess quality without relying on human raters, like we did by utilizing effort metrics like word count, could provide valuable insights.

Tables

Table 1

Reliability Metrics for Activity Quality and Efficiency

	κ	95% CI	Interpretation	Conditional Probability	
Quality					
Considers Problem	0.88	[0.63, 1.09]	Excellent	0.95	
Selects Problem	0.83	[0.62, 1.04]	Excellent	0.95	
Considers Practice	0.60	[0.39, 0.82]	Moderate	0.84	
Selects Practice	0.45	[0.24, 0.67]	Moderate	0.78	
	ICC			*F*	*p*
Efficiency	0.77	[0.64, 0.85]	Very Good	4.25	<0.001*

Table 2

Mixed-Effects Linear Regression Model with Total Words Predicting Total Minutes

		β	95% CI		S.E.	*t*	*df*	*p*
			Lower	Upper				
Total Minutes	Intercept	0.85	7.40	11.67	0.67	1.23	18.98	0.219
	Total Words	0.01	-0.00	-0.00	0.00	61.89	327.90	<0.001*

Table 3a

Descriptive Statistics for Primary Variables within CKS Condition and Comparison Group (Study 1)

	Mean	95% CI	*SD*	Median	Min	Max	IQR	Skewness	Kurtosis
Supervision 1									
Total Minutes	15.68	[11.85,19.51]	10.26	13.16	1.38	42.88	11.55	1.12	1.59
Weeks Post Training	46.84	[38.27, 55.41]	22.95	38.21	15.00	92.14	40.89	0.37	-1.19
Total Words	2256.73	[1671.84, 2841.63]	1566.37	1825.50	192.00	6669.00	2059.75	1.11	1.27
Considers Problem	920.17	[695.18, 1145.16]	602.53	804.00	0.00	2181.00	798.00	0.39	-0.51
Selects Problem	93.03	[33.92, 152.15]	158.32	38.50	0.00	670.00	114.25	2.65	7.28
Considers Practice	217.77	[132.29, 303.25]	228.91	160.50	0.00	681.00	381.00	0.79	-0.76
Selects Practice	109.13	[38.93, 179.34]	188.00	62.50	0.00	912.00	115.25	3.24	11.86
Efficiency	3.57	[3.05, 4.08]	1.38	4.00	1.00	5.00	2.00	-0.65	-0.67
Max Extensiveness									
Considers Problem	3.90	[3.29, 4.51]	1.63	5.00	0.00	5.00	2.00	-1.32	0.59
Selects Problem	3.40	[2.66, 4.14]	1.98	4.00	0.00	5.00	3.25	-0.95	-0.76
Considers Practice	2.70	[2.05, 3.35]	1.74	3.00	0.00	5.00	3.25	-0.51	-0.90
Selects Practice	2.90	[2.23, 3.57]	1.79	3.00	0.00	5.00	2.25	-0.58	-0.98
Supervision 2									
Total Minutes	13.82	[10.26,17.37]	9.17	11.30	1.18	35.35	10.82	0.98	0.28
Weeks Post Training	51.03	[42.36, 59.70]	22.36	46.36	19.00	92.43	43.39	0.34	-1.36
Total Words	2225.96	[1553.16, 2898.77]	1735.11	1798.00	137.00	6644.00	1882.75	1.31	1.24
Considers Problem	299.54	[112.41, 486.66]	482.58	118.00	0.00	2263.00	414.50	2.83	9.78

	Mean	95% CI	*SD*	Median	Min	Max	IQR	Skewness	Kurtosis
Selects Problem	20.75	[-1.01, 42.51]	56.12	0.00	0.00	232.00	0.00	3.18	9.70
Considers Practice	80.50	[39.05, 121.95]	106.90	0.00	0.00	332.00	160.25	0.94	-0.51
Selects Practice	76.54	[30.78, 122.29]	118.01	17.50	0.00	423.00	137.50	1.86	3.03
Efficiency	3.32	[2.85,3.79]	1.22	3.00	1.00	5.00	1.00	-0.28	-0.39
Max Extensiveness									
Considers Problem	2.00	[1.32,2.68]	1.76	2.00	0.00	5.00	3.00	0.26	-1.24
Selects Problem	1.11	[0.48,1.73]	1.62	0.00	0.00	5.00	2.75	1.23	0.35
Considers Practice	1.39	[0.87,1.91]	1.34	1.00	0.00	5.00	2.00	0.69	0.07
Selects Practice	1.71	[0.97,2.46]	1.92	1.00	0.00	5.00	3.00	0.61	-1.18
<u>Supervision 3</u>									
Total Minutes	9.78	[6.22,13.33]	8.80	6.66	1.10	33.72	7.53	1.77	2.46
Weeks Post Training	53.12	[44.49, 61.74]	21.37	52.21	27.86	93.29	42.00	0.36	-1.25
Total Words	1533.50	[975.79, 2091.21]	1380.78	1110.50	136.00	5442.00	1182.25	1.76	2.74
Considers Problem	185.31	[57.07, 313.54]	317.48	23.00	0.00	1305.00	269.25	2.29	5.54
Selects Problem	1.85	[-1.96, 5.64]	9.41	0.00	0.00	48.00	0.00	5.10	26.00
Considers Practice	52.81	[4.67, 100.96]	119.21	0.00	0.00	397.00	15.75	2.29	4.07
Selects Practice	38.81	[9.33, 68.29]	72.98	0.00	0.00	228.00	51.25	1.66	1.26
Efficiency	3.42	[2.86,3.98]	1.39	3.00	1.00	5.00	2.00	-0.46	-0.78
Max Extensiveness									
Considers Problem	1.19	[0.67,1.72]	1.30	1.00	0.00	3.00	3.00	0.45	-1.59
Selects Problem	0.73	[0.24,1.22]	1.22	0.00	0.00	3.00	1.25	1.28	-0.16

Considers Practice	0.69	[0.32,1.07]	0.93	0.00	0.00	3.00	1.25	1.01	-0.21
	Mean	95% CI	*SD*	Median	Min	Max	IQR	Skewness	Kurtosis
Selects Practice	1.04	[0.42,1.66]	1.54	0.00	0.00	5.00	3.00	1.08	-0.14
Total Supervision									
Total Minutes	13.23	[11.13, 15.33]	9.67	10.97	1.10	42.88	10.91	1.16	1.00
Weeks Post Training	50.18	[45.37, 54.99]	22.16	47.43	15.00	93.29	40.89	0.32	-1.23
Total Words	2022.62	[1678.15, 2367.09]	1587.32	1585.50	136.00	6669.00	1756.25	1.30	1.27
Considers Problem	485.83	[359.08, 612.59]	584.08	231.00	0.00	2263.00	793.25	1.33	1.11
Selects Problem	40.71	[17.54, 63.89]	106.80	0.00	0.00	670.00	35.00	4.15	19.87
Considers Practice	120.95	[82.31, 159.59]	178.05	0.00	0.00	681.00	197.75	1.58	1.74
Selects Practice	76.50	[46.34, 106.66]	138.97	4.00	0.00	912.00	111.75	3.50	16.28
Efficiency	3.44	[3.15, 3.73]	1.32	3.00	1.00	5.00	2.00	-0.45	-0.74
Max Extensiveness	1.85	[1.7, 1.99]	1.80	1.00	0.00	5.00	3.00	0.58	-1.02
Considers Problem	1.21	[1.07, 1.36]	1.85	0.00	0.00	5.00	3.00	1.13	-0.37
Selects Problem	1.12	[0.99, 1.24]	1.56	0.00	0.00	5.00	2.00	1.13	-0.06
Considers Practice	1.12	[0.99, 1.24]	1.56	0.00	0.00	5.00	2.00	1.13	-0.06
Selects Practice	1.20	[1.07, 1.34]	1.71	0.00	0.00	5.00	3.00	1.02	-0.47
Total Targets									
Considers Problem	1.88	[1.74, 2.01]	1.56	1.00	0.00	5.00	2.00	0.42	-1.15
Selects Problem	0.66	[0.58, 0.74]	0.93	0.00	0.00	5.00	1.00	1.43	1.96
Considers Practice	1.25	[1.10, 1.40]	1.80	0.00	0.00	9.00	2.00	1.45	1.30

Selects Practice	0.59	[0.52, 0.66]	0.85	0.00	0.00	4.00	1.00	1.55	2.44

Note. IQR = Interquartile Range. Total Targets indicates the total problems or practices. Supervision 1 had 30 total events, Supervision 2 had 28 total events, Supervision 3 had 26 total events, and Total Supervision had 84 total events.

Table 3b

Descriptive Statistics for Primary Variables within CKS Condition Only (Study 2)

	Mean	95% CI	*SD*	Median	Min	Max	IQR	Skewness	Kurtosis
Supervision 1									
Total Minutes	19.81	[18.00, 46.28]	9.84	18.91	2.80	46.28	13.88	0.483	-0.26
Weeks Post Training	57.81	[53.31, 62.32]	24.72	58.43	15.00	126.29	41.82	0.16	-0.62
Total Words	2859.87	[2560.50, 3159.25]	1642.07	2605.50	216.00	7503.00	2497.75	0.59	-0.37
Considers Problem	1247.92	[1106.11, 1389.74]	777.84	1181.50	63.00	4501.00	1050.25	0.98	2.00
Selects Problem	98.33	[77.07, 119.59]	116.63	64.00	0.00	670.00	137.75	2.31	7.07
Considers Practice	359.93	[288.35, 431.51]	392.62	226.50	0.00	2161.00	368.75	2.10	5.10
Selects Practice	132.11	[100.75, 163.47]	172.03	68.50	0.00	912.00	130.00	2.34	5.96
Efficiency	3.98	[3.82, 4.14]	0.88	4.00	1.00	5.00	1.00	-0.81	0.58
Max Extensiveness									
Considers Problem	4.56	[4.41, 4.71]	0.82	5.00	1.00	5.00	1.00	-1.74	2.38
Selects Problem	4.26	[4.07, 4.46]	1.08	5.00	0.00	5.00	1.00	-1.90	4.38
Considers Practice	3.14	[2.90, 3.39]	1.36	3.00	0.00	5.00	1.00	-0.68	0.18
Selects Practice	3.11	[2.80, 3.42]	1.69	3.00	0.00	5.00	2.00	-0.73	-0.62
Supervision 2									

Total Minutes	15.52	[13.83, 17.22]	12.75	8.93	4.47	47.80	12.68	1.17	1.11
Weeks Post Training	61.65	[57.05, 66.25]	24.45	61.00	18.57	126.57	43.57	0.14	-0.68
Total Words	2371.37	[2095.44, 2647.30]	1466.93	1929.00	64.00	7062.00	1830.00	1.24	1.41
Considers Problem	288.60	[226.38, 350.83]	330.82	202.00	0.00	1788.00	477.00	1.69	3.74

	Mean	95% CI	*SD*	Median	Min	Max	IQR	Skewness	Kurtosis
Selects Problem	27.05	[13.37, 40.74]	72.74	0.00	0.00	478.00	8.00	3.96	18.20
Considers Practice	209.78	[141.52, 278.05]	362.9	64.00	0.00	2302.00	241.00	3.06	11.70
Selects Practice	96.04	[72.41, 119.66]	125.62	41.00	0.00	511.00	146.00	1.56	1.76
Efficiency	3.67	[3.48, 3.8 7]	1.04	4.00	1.00	5.00	1.00	-0.50	-0.31
Max Extensiveness									
Considers Problem	2.06	[1.75, 2.38]	1.66	2.00	0.00	5.00	3.00	0.14	-1.21
Selects Problem	1.24	[0.94, 1.55]	1.63	0.00	0.00	5.00	3.00	0.96	-0.48
Considers Practice	1.59	[1.27, 1.91]	1.70	1.00	0.00	5.00	3.00	0.63	-0.87
Selects Practice	1.96	[1.64, 2.29]	1.75	2.00	0.00	5.00	3.00	0.22	-1.34
Supervision 3									
Total Minutes	10.80	[9.54, 12.07]	9.42	6.60	2.92	33.72	7.57	1.40	1.85
Weeks Post Training	64.94	[60.29, 69.60]	24.26	64.14	22.71	127.43	40.43	0.13	-0.67
Total Words	1734.17	[1532.28, 1936.06]	1053.36	1466.00	358.00	5151.00	1373.00	1.18	1.02
Considers Problem	197.19	[143.55, 250.82]	279.84	64.00	0.00	1305.00	315.00	1.90	3.86
Selects Problem	8.22	[0.60, 15.83]	39.71	0.00	0.00	323.00	0.00	6.20	42.53

Considers Practice	60.21	[26.06, 94.35]	178.15	0.00	0.00	1185.00	0.00	4.74	25.86
Selects Practice	32.78	[14.27, 51.28]	96.53	0.00	0.00	679.00	0.00	4.41	23.21
Efficiency	3.60	[3.40, 3.79]	1.01	4.00	1.00	5.00	1.00	-0.36	-0.26
Max Extensiveness									
Considers Problem	1.14	[0.91, 1.37]	1.18	1.00	0.00	3.00	2.00	0.47	-1.32
Selects Problem	0.56	[0.35, 0.77]	1.11	0.00	0.00	5.00	1.00	1.92	2.66

	Mean	95% CI	*SD*	Median	Min	Max	IQR	Skewness	Kurtosis
Considers Practice	0.50	[0.32, 0.69]	0.98	0.00	0.00	4.00	1.00	2.10	3.81
Selects Practice	0.55	[0.33, 0.77]	1.16	0.00	0.00	5.00	0.00	1.98	2.82
<u>Total Supervision</u>									
Total Minutes	15.38	[14.38, 16.38]	9.33	12.73	2.80	47.80	12.94	0.99	0.48
Weeks Post Training	61.39	[58.77, 64.00]	24.55	61.29	15.00	127.43	42.21	0.13	-0.66
Total Words	2322.79	[2164.70, 2480.89]	1484.22	1924.00	64.00	7503.00	1934.00	1.08	0.73
Considers Problem	588.65	[513.29, 664.02]	707.58	313.00	0.00	4501.00	850.00	1.73	3.77
Selects Problem	45.41	[35.63, 55.19]	91.82	0.00	0.00	670.00	56.00	3.21	13.13
Considers Practice	213.05	[175.99, 250.12]	347.99	85.00	0.00	2302.00	266.00	2.77	9.40
Selects Practice	87.45	[72.41, 102.50]	141.22	25.00	0.00	912.00	123.50	2.58	7.98
Efficiency	3.75	[3.64, 3.85]	0.98	4.00	1.00	5.00	1 .00	-0.55	-0.13
Max Extensiveness	2.61	[2.41, 2.82]	1.93	3.00	0.00	5.00	4.00	-0.09	-1.47
Considers Problem	2.06	[1.83, 2.28	2.08	1.00	0.00	5.00	4.00	0.28	-1.62
Selects Problem	1.78	[1.60, 1.96]	1.76	2.00	0.00	5.00	3.00	0.40	-1.25

Considers Practice	1.90	[1.70, 2.10]	1.88	2.00	0.00	5.00	3.50	0.30	-1.46
Selects Practice	2.61	[2.41, 2.82]	1.93	3.00	0.00	5.00	4.00	-0.09	-1.47
Total Targets									
Considers Problem	2.61	[2.41, 2.82]	1.93	3.00	0.00	5.00	4.00	-0.09	-1.47
Selects Problem	2.06	[1.83, 2.28]	2.08	1.00	0.00	5.00	4.00	0.28	-1.62
Considers Practice	1.78	[1.60, 1.96]	1.76	2.00	0.00	5.00	3.00	0.40	-1.25
Selects Practice	1.90	[1.70, 2.10]	1.88	2.00	0.00	5.00	3.50	0.30	-1.46

Note. IQR = Interquartile Range. Total Targets indicates the total problems or practices. Supervision 1 has 30 events, Supervision 2 has 28 events, Supervision 3 has 26 events, and Total Supervision has 84 events.

Table 4a

Frequency of Quality and Effort in Supervision Events per Activity within CKS and Comparison Group condition (Study 1)

	First Supervision (n = 30)		Second Supervision (n = 28)		Third Supervision (n = 26)		Total (N = 84)	
	Count	%tage	Count	%tage	Count	%tage	Count	%tage
Quality								
Considers Problem								
Absent	6.00	20.00	22.00	78.60	22.00	84.60	50.00	59.50
Present	24.00	80.00	6.00	21.40	4.00	15.40	34.00	40.50
Selects Problem								
Absent	11.00	36.70	22.00	78.60	26.00	100.00	59.00	70.20
Present	19.00	63.30	6.00	21.40	0.00	0.00	25.00	29.80
Considers Practice								
Absent	9.00	30.00	20.00	71.40	23.00	88.50	52.00	61.90
Present	21.00	70.00	8.00	28.60	3.00	11.50	32.00	38.10
Selects Practice								
Absent	15.00	50.00	20.00	71.40	19.00	73.10	54.00	64.30
Present	15.00	50.00	8.00	28.60	7.00	26.90	30.00	35.70
Effort								
Considers Problem								
Absent	2.00	6.70	8.00	28.60	11.00	42.30	21.00	25.00
Present	28.00	93.30	20.00	71.40	15.00	57.70	63.00	75.00
Selects Problem								
Absent	12.00	40.00	20.00	71.40	24.00	92.30	56.00	66.70
Present	18.00	60.00	8.00	28.60	2.00	7.70	28.00	33.30
Considers Practice								
Absent	7.00	23.30	11.00	39.30	16.00	61.50	34.00	40.50
Present	23.00	76.70	17.00	60.70	10.00	38.50	50.00	59.50
Selects Practice								
Absent	6.00	20.00	12.00	42.90	17.00	65.40	35.00	41.70
Present	24.00	80.00	16.00	57.10	9.00	34.60	49.00	58.30

Note. Presence of effort denotes that words were spoken within that activity.

Table 4b

Frequency of Quality and Effort in Supervision Events per Activity within CKS condition (Study 2)

	First Supervision (n = 118)		Second Supervision (n = 111)		Third Supervision (n = 107)		Total (N = 341)	
	Count	%tage	Count	%tage	Count	%tage	Count	%tage
Quality								
Considers Problem								
Absent	4.00	3.40	77.00	69.40	92.00	86.00	178.00	52.20
Present	114.00	96.60	34.00	30.60	15.00	14.00	163.00	47.80
Selects Problem								
Absent	25.00	21.20	82.00	73.90	101.00	94.40	213.00	62.50
Present	93.00	78.80	29.00	26.10	6.00	5.60	128.00	37.50
Considers Practice								
Absent	14.00	11.90	63.00	56.80	94.00	87.90	175.00	51.30
Present	104.00	88.10	48.00	43.20	13.00	12.10	166.00	48.70
Selects Practice								
Absent	31.00	26.30	56.00	50.50	90.00	84.10	181.00	53.10
Present	87.00	73.70	55.00	49.50	17.00	15.90	160.00	46.90
Effort								
Considers Problem								
Absent	2.00	1.70	30.00	27.00	47.00	43.90	82.00	24.00
Present	116.00	98.30	81.00	73.00	60.00	56.10	259.00	76.00
Selects Problem								
Absent	24.00	20.30	82.00	73.90	99.00	92.50	210.00	61.60
Present	94.00	79.70	29.00	26.10	8.00	7.50	131.00	38.40
Considers Practice								
Absent	10.00	8.50	47.00	42.30	83.00	77.60	144.00	42.20
Present	108.00	91.50	64.00	57.70	24.00	22.40	197.00	57.80
Selects Practice								
Absent	18.00	15.30	38.00	34.20	87.00	81.30	147.00	43.10
Present	100.00	84.70	73.00	65.80	20.00	18.70	194.00	56.90

Note. Presence of effort denotes that words were spoken within that activity.

Table 5

Total Cases and Events per Supervisor and Supervisee; Total Supervisees per Supervisor

	Mean	95% CI	*SD*	Median	Min	Max	IQR	Skewness	Kurtosis
Study 1									
Supervisors (*n* = 17)									
Total Events	23.00	[17.99, 28.01]	9.75	21.00	5.00	39.00	18.00	0.00	-0.94
Total Supervisees	3.53	[2.80, 4.26]	1.42	4.00	2.00	6.00	3.00	0.38	-1.03
Total Cases	7.94	[6.24, 9.64]	3.31	7.00	3.00	13.00	7.00	0.13	-1.49
Supervisees (*n* = 26)									
Total Events	7.50	[6.15, 8.85]	3.34	7.50	2.00	14.00	3.00	0.07	-0.61
Total Cases	2.58	[2.18, 2.98]	0.99	3.00	1.00	4.00	1.00	-0.10	-0.36
Study 2									
Supervisors *(n* = 16)									
Total Events	21.31	[15.32, 27.31]	11.25	21.00	3.00	39.00	18.00	0.06	-0.89
Total Supervisees	3.06	[2.27, 3.85]	1.48	3.00	1.00	6.00	2.00	0.44	-0.66
Total Cases	7.31	[5.29, 9.33]	3.79	7.00	1.00	12.00	8.00	0.19	-1.08
Supervisees (*n* = 48)									
Total Events	7.10	[6.18, 8.03]	3.19	7.00	2.00	14.00	5.00	0.23	-0.82
Total Cases	2.44	[2.14, 2.74]	1.03	2.50	1.00	4.00	1.00	0.17	-0.76

Note. IQR = Interquartile Range.

Table 6

Firth-corrected Logistic Regression Predicting the Likelihood of Activity Quality with Supervision Type, Repetition of Supervision Activities, and Weeks Post Training as Covariates

	β	*SE* β	Wald's X^2	OR	95% CI	*df*	*p*
Considers Problem							
Intercept	4.34	0.60	Inf	76.96	[23.77, 249.19]	3.00	**<0.01***
Supervision Type	-2.33	0.24	Inf	0.10	[0.06, 0.16]	3.00	**<0.01***
Repetition of Activities	-0.03	0.06	0.25	0.97	[0.86, 1.10]	3.00	0.62
Weeks Post Training	0.00	0.00	0.32	1.00	[0.99, 1.02]	3.00	0.57
Selects Problem							
Intercept	3.69	0.56	54.42	40.38	[13.42, 121.50]	3.00	**<0.01***
Supervision Type	-2.11	0.23	Inf	0.12	[0.08, 0.19]	3.00	**<0.01***
Repetition of Activities	-0.01	0.06	0.05	0.99	[0.87, 1.11]	3.00	0.83
Weeks Post Training	0.00	0.00	0.51	1.00	[0.98, 1.01]	3.00	0.47
Considers Practice							
Intercept	4.83	0.60	Inf	20.86	[8.09, 53.79]	3.00	**<0.01***
Supervision Type	-1.91	0.21	Inf	0.29	[0.21, 0.40]	3.00	**<0.01***
Repetition of Activities	0.00	0.06	0.01	0.98	[0.88, 1.09]	3.00	0.70
Weeks Post Training	-0.02	0.01	6.43	0.99	[0.98, 1.00]	3.00	0.10
Selects Practice							
Intercept	3.04	0.48	46.83	124.8	[38.32, 406.37]	3.00	**<0.01***
Supervision Type	-1.23	0.17	64.66	0.15	[0.10, 0.22]	3.00	**<0.01***
Repetition of Activities	-0.02	0.05	0.15	1.00	[0.89, 1.13]	3.00	0.94
Weeks Post Training	-0.01	0.01	2.67	0.98	[0.97, 1.00]	3.00	**0.01***

Note. OR = Odds Ratio. The p-value cutoff for statistical significance was set at $p < 0.05$. The p-value is for the slope coefficient. The analysis for Selects Practice was conducted with and without the transformed Weeks Post Training variable yielding comparable results. The table presents the analysis results using the non-transformed Weeks Post Training. Model fitted by Penalized ML.

Table 7a

Firth-corrected Logistic Regression Predicting the Likelihood of Activity Quality: First Supervision

	β	*SE* β	Wald's X^2	OR	95% CI	*df*	*p*
Considers Problem							
Intercept	4.07	1.40	11.31	58.27	[3.72, 913.97]	2.00	**<0.01***
Case	-0.04	0.17	0.05	0.96	[0.69, 1.33]	2.00	0.822
Weeks Post Training	-0.01	0.02	0.28	0.99	[0.95, 1.03]	2.00	0.598
Selects Problem							
Intercept	1.68	0.65	7.41	5.37	[1.51,19.13]	2.00	**<0.02***
Case	-0.04	0.09	0.22	0.96	[0.80, 1.14]	2.00	0.642
Weeks Post Training	0.00	0.01	0.14	1.00	[0.98, 1.02]	2.00	0.712
Considers Practice							
Intercept	1.22	0.59	4.45	3.39	[1.06, 10.86]	2.00	**0.04***
Case	0.04	0.08	0.17	1.04	[0.87, 1.23]	2.00	0.679
Weeks Post Training	-0.05	0.01	0.28	1.00	[0.98, 1.01]	2.00	0.599
Selects Practice							
Intercept	2.88	0.86	13.99	17.83	[3.29, 96.62]	2.00	**<0.01***
Case	0.04	0.11	0.10	1.04	[0.84, 1.29]	2.00	0.747
Weeks Post Training	-0.01	0.01	1.42	0.99	[0.96, 1.01]	2.00	0.234

Note. OR = Odds Ratio. The p-value cutoff for statistical significance was set at $p < 0.05$. The p-value is for the slope coefficient. Degrees of freedom (*df*) = 2. The analysis for Selects Practice was conducted with and without the transformed Weeks Post Training variable yielding comparable results. The table presents the analysis results using the non-transformed Weeks Post Training.

Table 7b

Firth-corrected Logistic Regression Predicting the Likelihood of Activity Quality: Second Supervision

	β	*SE* β	Wald's X^2	OR	95% CI	*df*	*p*
Considers Problem							
Intercept	-0.52	0.58	0.82	0.60	[0.19, 1.84]	2.00	0.39
Case	-0.01	0.09	0.01	0.99	[0.83, 1.18]	2.00	0.92
Weeks Post Training	-0.03	0.01	0.13	1.00	[0.98, 1.01]	2.00	0.72
Selects Problem							
Intercept	-0.72	0.60	7.41	0.49	[0.15, 1.58]	2.00	0.23
Case	0.02	0.09	0.22	1.02	[0.85, 1.22]	2.00	0.85
Weeks Post Training	-0.01	0.01	0.14	0.99	[0.98, 1.01]	2.00	0.58
Considers Practice							
Intercept	0.36	0.54	1.46	1.44	[0.49, 4.13]	2.00	0.51
Case	-0.02	0.08	0.04	0.14	[0.84, 1.15]	2.00	0.85
Weeks Post Training	-0.004	0.01	0.30	0.01	[0.98, 1.01]	2.00	0.61
Selects Practice							
Intercept	0.61	0.55	1.24	1.84	[0.62, 5.45]	2.00	0.27
Case	0.08	0.08	0.91	1.08	[0.92, 1.28]	2.00	0.34
Weeks Post Training	-0.02	0.01	4.08	0.98	[0.97, 1.00]	2.00	**0.04***

Note. OR = Odds Ratio. The p-value cutoff for statistical significance was set at $p < 0.05$. The p-value is for the slope coefficient. Degrees of freedom (*df*) = 2. The analysis for Selects Practice was conducted with and without the transformed Weeks Post Training variable yielding comparable results. The table presents the analysis results using the non-transformed Weeks Post Training.

Table 7c

Firth-corrected Logistic Regression Predicting the Likelihood of Activity Quality: Third Supervision

	β	*SE* β	Wald's X^2	OR	95% CI	*df*	*p*
Considers Problem							
Intercept	-3.10	0.97	12.91	0.04	[0.01, 0.30]	2.00	**<0.01***
Case	-0.05	0.10	0.24	0.95	[0.78, 1.16]	2.00	0.63
Weeks Post Training	0.20	0.01	2.95	1.02	[1.00, 1.04]	2.00	0.09
Selects Problem							
Intercept	-2.19	1.11	4.02	0.11	[0.01, 1.00]	2.00	0.05
Case	-0.01	0.17	0.00	0.99	[0.72, 1.37]	2.00	0.97
Weeks Post Training	-0.01	0.02	0.08	0.99	[0.96, 1.03]	2.00	0.77
Considers Practice							
Intercept	0.60	0.78	0.57	1.81	[0.39, 8.38]	2.00	0.45
Case	-0.19	0.14	1.84	0.83	[0.63, 1.09]	2.00	0.18
Weeks Post Training	-0.02	0.01	2.36	0.98	[0.96, 1.01]	2.00	0.12
Selects Practice							
Intercept	0.02	0.87	0.00	1.02	[0.19, 5.65]	2.00	0.98
Case	-0.29	0.16	3.49	0.75	[0.55, 1.03]	2.00	0.06
Weeks Post Training	-0.01	0.01	0.29	0.99	[0.97, 1.02]	2.00	0.59

Note. OR = Odds Ratio. The p-value cutoff for statistical significance was set at $p < 0.05$. The p-value is for the slope coefficient. Degrees of freedom (*df*) = 2. The analysis for Selects Practice was conducted with and without the transformed Weeks Post Training variable yielding comparable results. The table presents the analysis results using the non-transformed Weeks Post Training.

Table 8a

Mixed-Effects Linear Regression Model Predicting Efficiency

	β	95% CI	*SE β*	*t*-value	*df*	*p*
First Supervision						
Intercept	3.66	[3.14, 4.17]	0.26	14.01	59.80	**<0.01***
Case	0.00	[-0.07, 0.07]	0.00	0.06	109.90	0.95
Weeks Post Training	0.00	[-0.00, 0.01]	0.00	1.17	69.84	0.25
Second Supervision						
Intercept	3.39	[2.78, 4.00]	0.31	10.58	67.67	**<0.01***
Case	0.02	[-0.07, 0.10]	0.00	0.07	103.80	0.70
Weeks Post Training	0.00	[-0.01, 0.01]	0.00	0.22	76.30	0.55
Third Supervision						
Intercept	3.51	[2.85, 4.17]	0.31	10.58	67.67	**<0.01***
Case	0.00	[-0.08, 0.08]	0.00	0.07	103.80	0.96
Weeks Post Training	0.00	[-0.01, 0.01]	0.00	0.22	76.30	0.83

Table 9

Mixed-Effects Linear Regression Model Predicting Efficiency with Supervision Type, Repetition of Supervision Activities, and Weeks Post Training as Covariates

	β	95% CI	*SE β*	*t*-value	*df*	*p*
Intercept	3.88	[3.39, 4.37]	0.25	15.51	74.16	**<0.01***
Supervision Type	-0.19	[-0.31, -0.08]	0.03	-0.30	139.26	**0.01***
Repetition of Activities	-0.01	[-0.07, 0.05]	0.00	1.23	90.94	0.77
Weeks Post Training	0.00	[-0.00, 0.01]	0.06	-3.40	325.22	0.22

Table 10

The Likelihood of Activity Effort based on Case and Weeks Post Training with Supervision Type, Repetition of Supervision Activities, and Weeks Post Training as Covariates

Considers Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	1924.41	[1365.04, 2712.99]	7.56	0.18	43.16	**<0.001**
Supervision Type	0.50	[0.45, 0.56]	-0.69	0.06	-11.79	**<0.001**
Repetition of Activities	0.98	[0.94, 1.02]	-0.02	0.02	-0.92	0.357
Weeks Post Training	1.00	[1.00, 1.01]	0.00	0.00	1.11	0.268
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.01	[0.00, 0.04]	-4.84	0.82	-5.91	**<0.001**
Supervision Type	4.88	[2.86, 8.32]	1.56	0.27	5.83	**<0.001**
Repetition of Activities	1.14	[0.96, 1.36]	0.13	0.09	1.50	0.134
Weeks Post Training	0.99	[0.97, 1.01]	-0.01	0.01	-1.26	0.208
Selects Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	119.72	[67.72, 211.65]	4.79	0.29	16.46	**<0.001**
Supervision Type	0.94	[0.71, 1.24]	-0.06	0.14	-0.44	0.658
Repetition of Activities	0.95	[0.88, 1.03]	-0.05	0.04	-1.31	0.190
Weeks Post Training	1.00	[0.99, 1.01]	0.00	0.00	0.57	0.568
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.01	[0.00, 0.04]	-4.58	0.70	-6.58	**<0.001**
Supervision Type	10.74	[6.14, 18.77]	2.37	0.28	8.33	**<0.001**
Repetition of Activities	1.04	[0.89, 1.22]	0.04	0.08	0.51	0.612
Weeks Post Training	1.01	[0.99, 1.03]	0.01	0.01	0.97	0.333
Considers Practice						

Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	453.27	[269.80, 761.51]	6.12	0.26	23.12	**<0.001**
Supervision Type	0.87	[0.73, 1.04]	-0.14	0.10	-1.52	0.129
Repetition of Activities	0.95	[0.89, 1.02]	-0.05	0.04	-1.47	0.142
Weeks Post Training	1.00	[0.99, 1.01]	0.00	0.00	0.03	0.973
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.01	[0.00, 0.03]	-4.84	0.72	-6.84	**<0.001**
Supervision Type	6.07	[3.75, 9.83]	1.80	0.25	7.34	**<0.001**
Repetition of Activities	1.06	[0.93, 1.22]	0.06	0.07	0.87	0.385
Weeks Post Training	1.01	[0.99, 1.02]	0.01	0.01	0.88	0.377
Selects Practice						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	109.88	[67.21, 179.64]	4.70	0.25	18.74	**<0.001**
Supervision Type	1.08	[0.89, 1.31]	0.08	0.10	0.78	0.436
Repetition of Activities	0.91	[0.85, 0.98]	-0.09	0.04	-2.36	**0.018**
Weeks Post Training	1.01	[1.00, 1.01]	0.00	0.00	1.57	0.117
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.01	[0.00, 0.04]	-4.61	0.71	-6.53	**<0.001**
Supervision Type	6.35	[3.93, 10.26]	1.85	0.25	7.55	**<0.001**
Repetition of Activities	1.08	[0.93, 1.26]	0.08	0.08	1.04	0.298
Weeks Post Training	1.00	[0.98, 1.02]	0.00	0.01	0.10	0.918

Note. IRR = Incidence Rate Ratios, which are exponentiated coefficients. OR = Odds Ratios, which are also exponentiated coefficients.

Table 11a

The Likelihood of Activity Effort based on Repetition of Case and Weeks Post Training: First Supervision

Considers Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	1395.36	[931.77, 2089.61]	7.24	0.20	35.14	**<0.001**
Case	1.07	[1.02, 1.12]	0.06	0.02	2.81	**0.005**
Weeks Post Training	0.99	[0.99, 1.00]	-0.01	0.00	-3.13	**0.002**
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.10	[0.00, Inf]	-2.26	81566.10	0.00	1.00
Case	0.33	[0.00, Inf]	-1.11	31224.42	0.00	1.00
Weeks Post Training	0.39	[0.00, Inf]	-0.94	5762.71	0.00	1.00
Selects Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	108.06	[65.44, 178.44]	4.68	0.26	18.30	**<0.001**
Case	0.95	[0.87, 1.04]	-0.05	0.05	-1.17	0.242
Weeks Post Training	1.00	[0.99, 1.01]	0.00	0.00	0.60	0.551
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.00	[0.00, 0.00]	-12.60	3.56	-3.55	**<0.001**
Case	1.02	[0.38, 2.76]	0.02	0.51	0.05	0.962
Weeks Post Training	1.01	[0.90, 1.13]	0.01	0.06	0.17	0.865
Considers Practice						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	307.5	[187.93, 503.14]	5.73	0.30	23.80	**<0.001**
Case	0.96	[0.88, 1.04]	-0.04	0.04	-1.00	0.319
Weeks Post Training	1.00	[0.99, 1.01]	0.01	0.01	0.78	0.434
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*

Intercept	0	[0.00, 0.09]	-13.87	5.83	-2.38	**0.017**
Case	1.03	[0.21, 5.15]	0.03	0.82	-0.04	0.972
Weeks Post Training	1.01	[0.83, 1.22]	0.01	0.10	0.10	0.925
Selects Practice Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	156.18	[91.45, 266.71]	5.05	0.27	18.50	**<0.001**
Case	0.91	[0.82, 1.00]	-0.10	0.05	-1.97	**0.049**
Weeks Post Training	1.00	[0.99, 1.01]	0.00	0.01	0.10	0.891
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.00	[0.00, 0.02]	-11.04	3.71	-2.98	**0.003**
Case	1.18	[0.27, 5.22]	0.16	0.76	0.22	0.83
Weeks Post Training	0.97	[0.81, 1.15]	-0.04	0.09	-0.40	0.69

Note. IRR = Incidence Rate Ratios, which are exponentiated coefficients. OR = Odds Ratios, which are also exponentiated coefficients.

Inf = infinity.

Table 11b

The Likelihood of Activity Effort based on Case and Weeks Post Training: Second Supervision

Considers Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	225.66	[131.81, 386.35]	5.42	0.27	19.75	**<0.001**
Case	1.00	[0.93, 1.08]	0.00	0.04	0.01	0.989
Weeks Post Training	1.01	[1.00, 1.02]	0.01	0.00	1.47	0.142
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.35	[0.10, 1.27]	-1.04	0.66	-1.59	0.111
Case	1.06	[0.85, 1.32]	0.06	0.11	0.55	0.583
Weeks Post Training	0.99	[0.97, 1.02]	0.00	0.01	0.40	0.692
Selects Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	98.6	[23.67, 410.77]	4.59	0.73	6.31	**<0.001**
Case	1.00	[0.79, 1.26]	0.00	0.12	-0.02	0.978
Weeks Post Training	1.00	[0.97, 1.03]	0.00	0.01	-0.17	0.863
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	2.36	[0.35, 15.92]	0.86	0.97	0.88	0.378
Case	0.97	[0.73, 1.30]	-0.03	0.15	-0.19	0.852
Weeks Post Training	1.02	[0.98, 1.06]	0.02	0.02	0.80	0.423
Considers Practice						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	433.55	[210.91, 891.24]	6.07	0.37	16.52	**<0.001**
Case	0.94	[0.85, 1.05]	-0.06	0.05	-1.05	0.294
Weeks Post Training	1.00	[0.98, 1.01]	0.00	0.01	0.40	0.555
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.5	[0.15, 1.64]	-0.70	0.61	-1.15	0.251
Case	1.01	[0.83, 1.23]	0.01	0.09	0.14	0.889
Weeks Post Training	1.00	[0.98, 1.03]	0.00	0.01	0.40	0.691
Selects Practice						

Count Model	IRR	95% CI	β	*SE* β	*z*-value	*p*
Intercept	95.42	[44.52, 204.52]	0.38	0.39	11.72	**<0.001**
Case	0.94	[0.84, 1.04]	0.06	0.06	-1.20	0.231
Weeks Post Training	1.01	[1.00, 1.02]	0.01	0.01	1.63	0.104
Zero-Inflated Model	OR	95% CI	β	*SE* β	*z*-value	*p*
Intercept	0.15	[0.03, 0.88]	-1.90	0.89	-2.10	**0.036**
Case	0.98	[0.76, 1.27]	-0.02	0.13	-0.13	0.894
Weeks Post Training	1.01	[0.98, 1.05]	0.01	0.02	0.76	0.449

Note. IRR = Incidence Rate Ratios, which are exponentiated coefficients. OR = Odds Ratios, which are also exponentiated coefficients.

Table 11c

The Likelihood of Activity Effort based on Case and Weeks Post Training: Third Supervision

Considers Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	211.04	[111.26, 400.32]	5.35	0.32	16.39	**<0.001**
Case	0.97	[0.89, 1.06]	-0.03	0.04	-0.72	0.474
Weeks Post Training	1.01	[1.00, 1.02]	0.00	0.01	1.67	0.097
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.85	[0.26, 2.86]	-0.15	0.62	-0.26	0.798
Case	1.14	[0.93, 1.40]	0.14	0.10	1.31	0.192
Weeks Post Training	0.99	[0.96, 1.01]	-0.01	0.01	-1.12	0.261
Selects Problem						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	29.09	[10.44, 81.04]	3.37	0.52	6.46	**<0.001**
Case	0.80	[0.77, 0.83]	-0.22	0.02	-11.84	**<0.001**
Weeks Post Training	1.03	[1.02, 1.04]	0.03	0.01	5.26	**<0.001**
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	5.94	[0.24, 145.53]	1.78	1.63	1.10	0.275
Case	1.09	[0.71, 1.68]	0.09	0.22	0.40	0.686
Weeks Post Training	1.02	[0.96, 1.07]	0.02	0.03	0.59	0.557
Considers Practice						
Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	380.75	[210.63, 688.27]	5.94	0.30	19.67	**<0.001**
Case	1.07	[1.02, 1.12]	0.06	0.02	2.75	**0.006**
Weeks Post Training	0.98	[0.97, 0.99]	-0.20	0.01	-3.34	**0.001**

Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.89	[0.23, 3.45]	-0.12	0.69	-0.17	0.862
Case	1.18	[0.94, 1.50]	0.17	0.12	1.41	0.158
Weeks Post Training	1.01	[0.98, 1.03]	0.01	0.01	0.58	0.559
Selects Practice Count Model	IRR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	136.61	[63.33, 294.67]	4.91	0.39	12.54	**<0.001**
Case	0.86	[0.68, 1.09]	-0.15	0.12	-1.28	0.2
Weeks Post Training	1.02	[0.99, 1.04]	0.01	0.01	1.23	0.219
Zero-Inflated Model	OR	95% CI	*β*	*SE β*	*z*-value	*p*
Intercept	0.75	[0.17, 3.25]	-0.23	0.75	-0.39	0.700
Case	1.19	[0.91, 1.56]	0.18	0.14	1.29	0.198
Weeks Post Training	1.01	[0.99, 1.04]	0.01	0.01	1.03	0.305

Note. IRR = Incidence Rate Ratios, which are exponentiated coefficients. OR = Odds Ratios, which are also exponentiated coefficients.

Figures

Figure 1a

First Supervision: Quality Across Cases

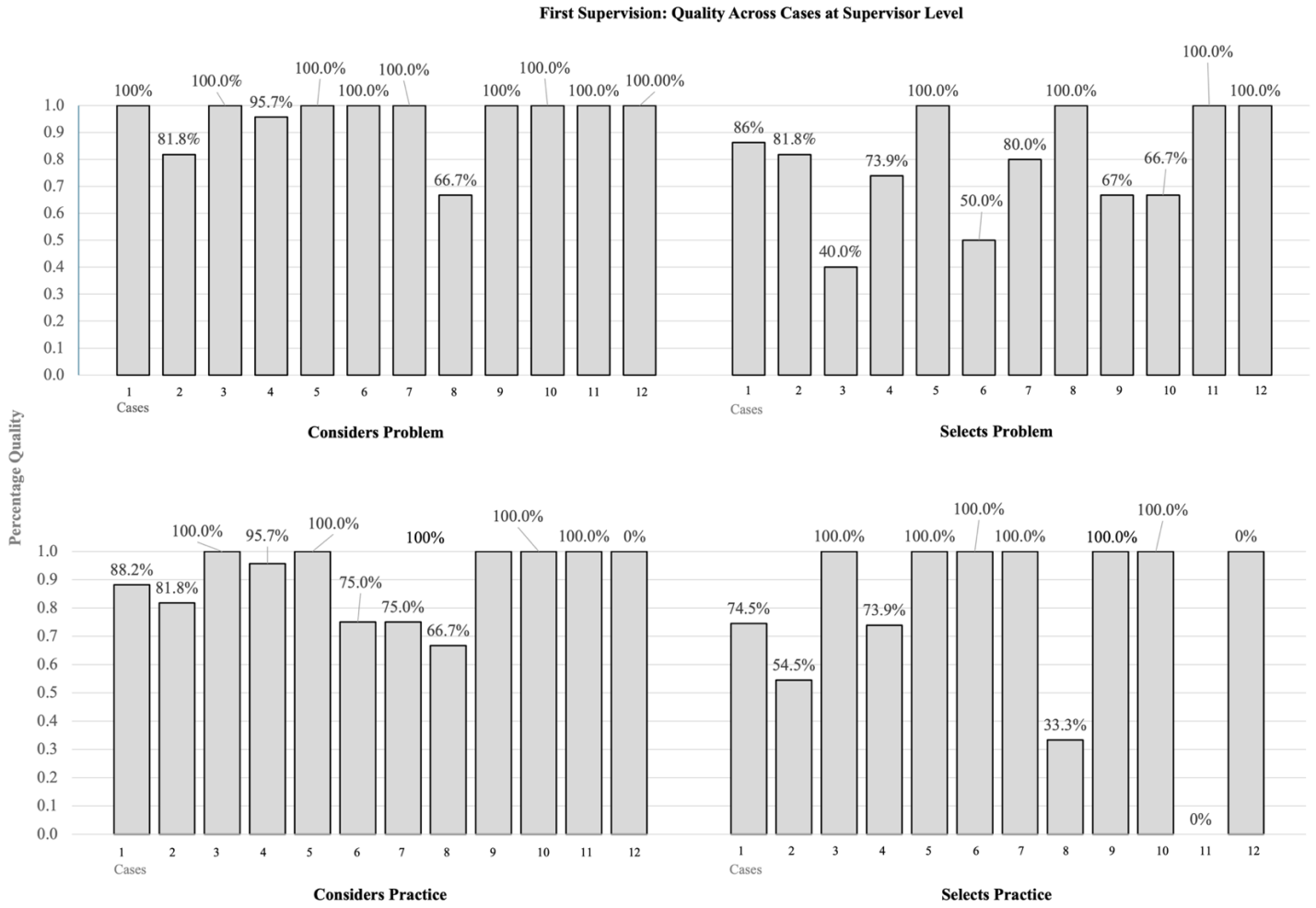


Figure 1b

Second Supervision: Quality Across Cases

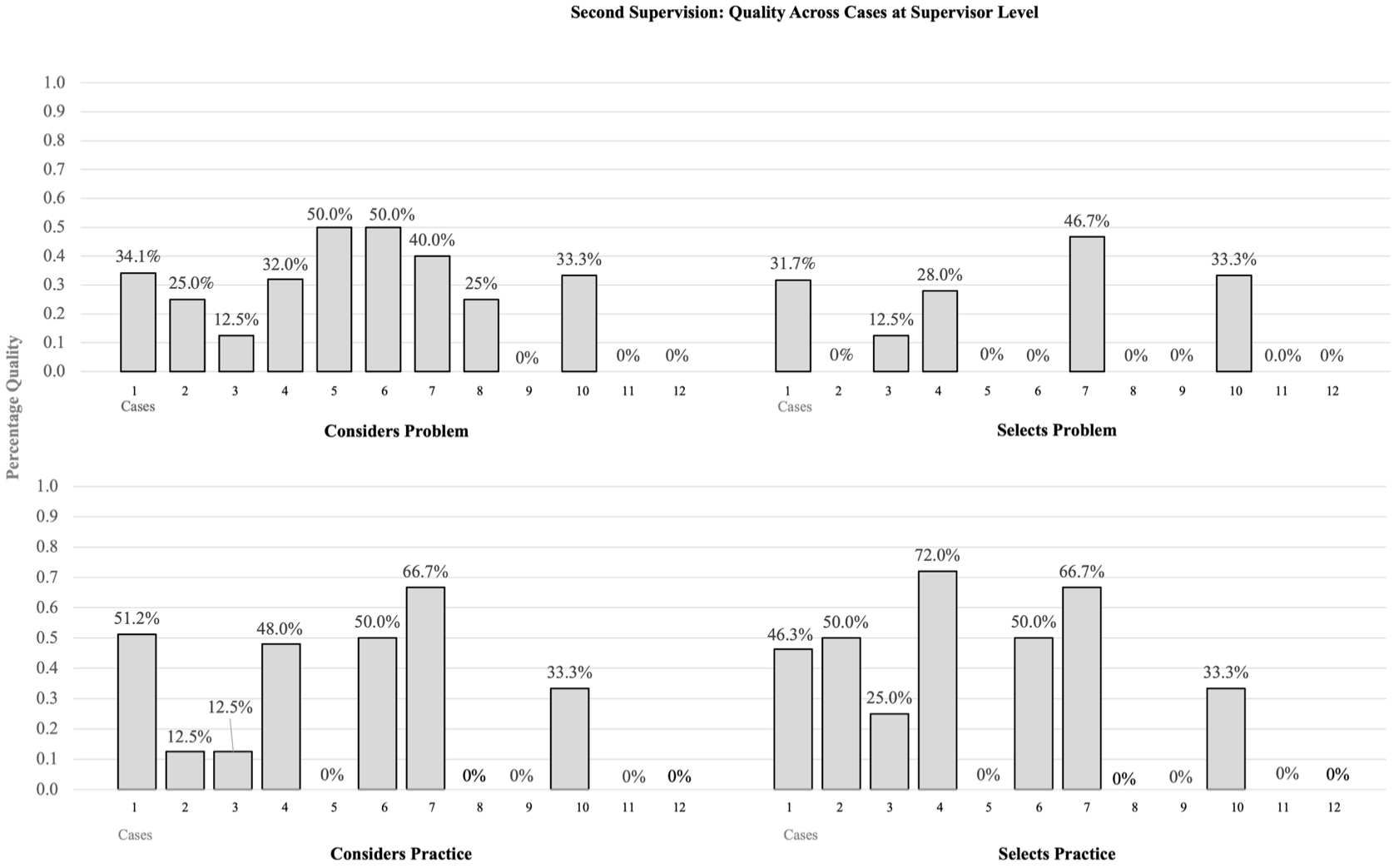


Figure 1c

Third Supervision: Quality Across Cases

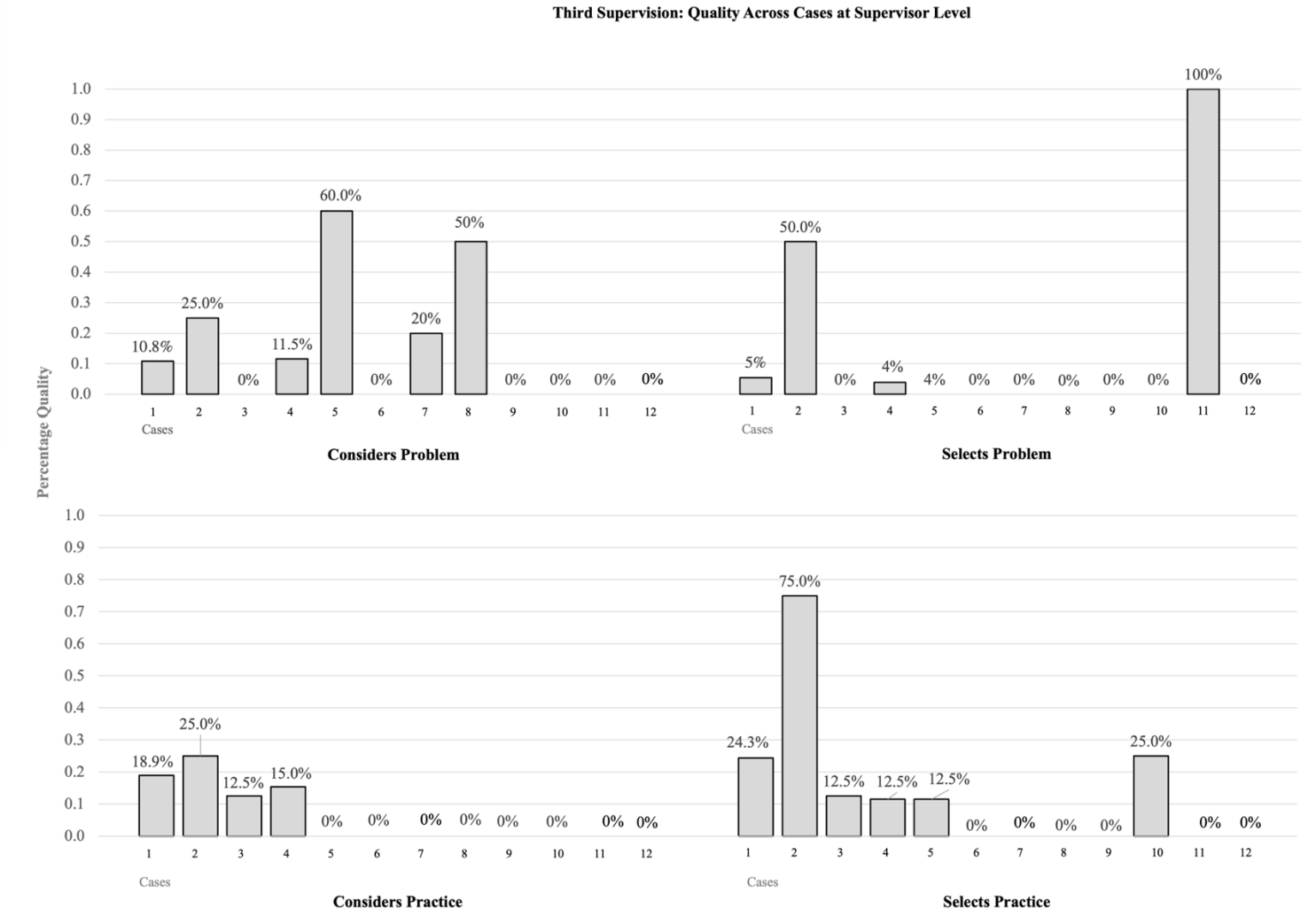


Figure 2a

First Supervision: Effort Across Cases

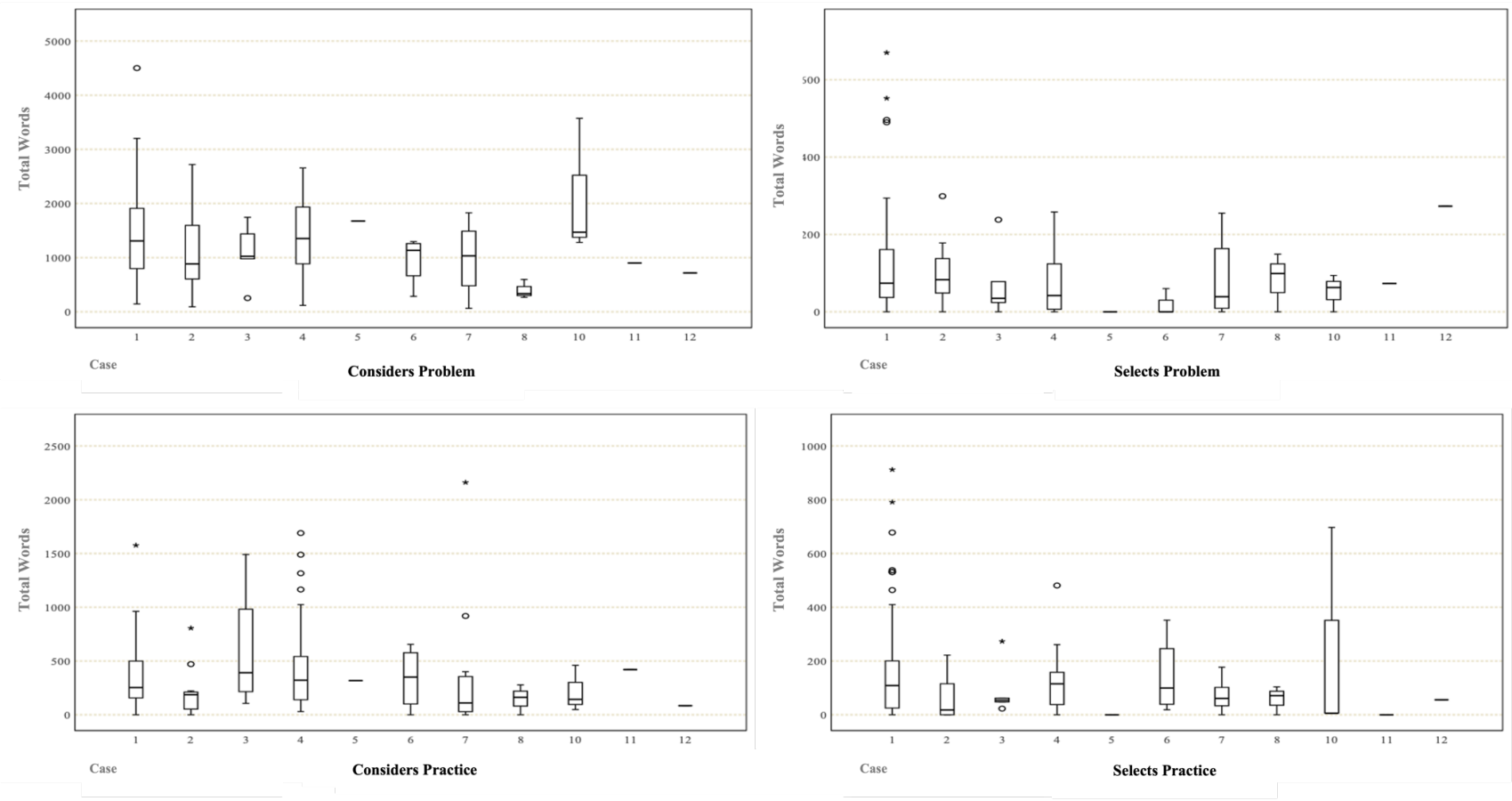


Figure 2b

Second Supervision: Effort Across Cases

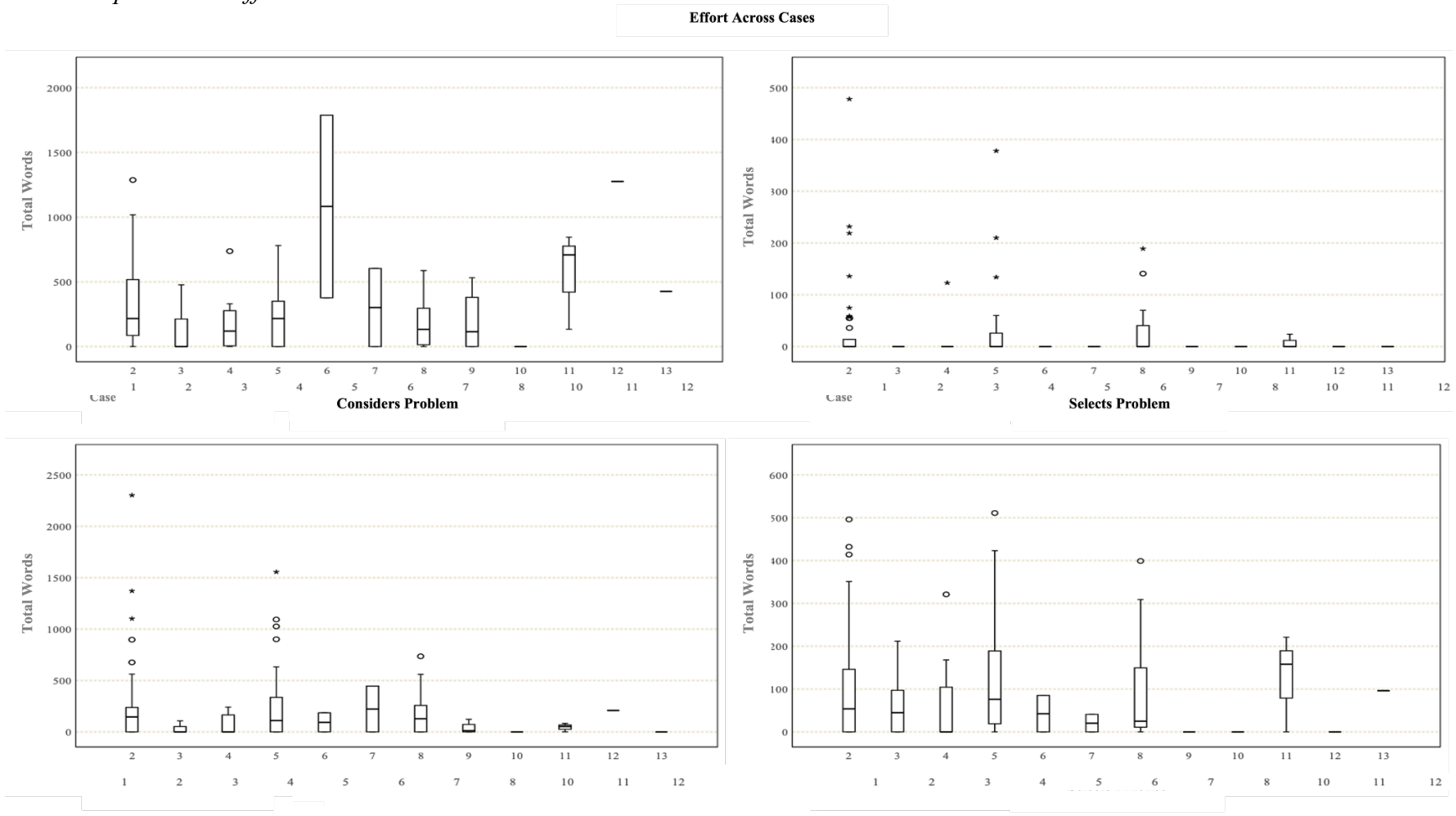


Figure 2c

Third Supervision: Effort Across Cases

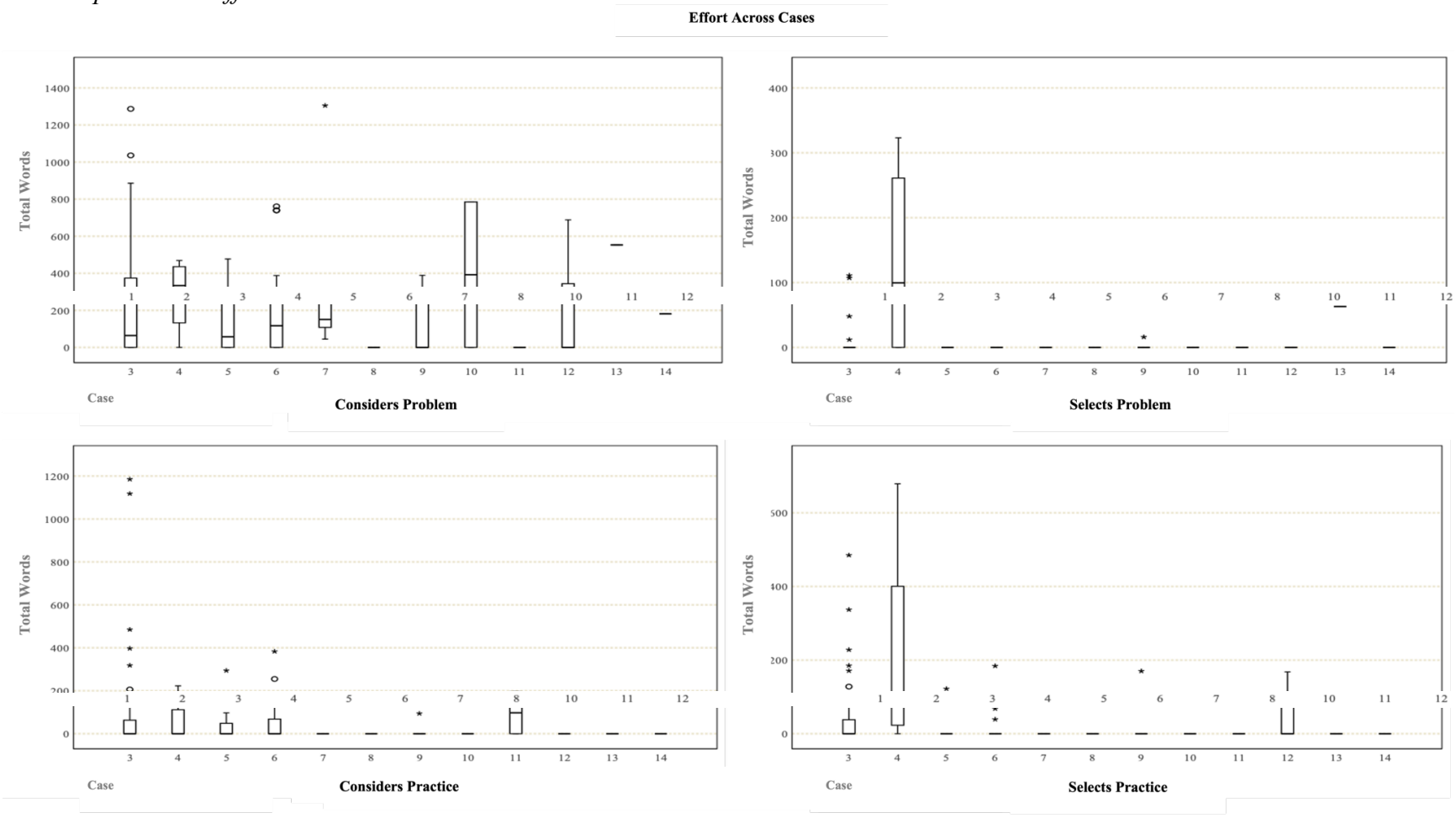
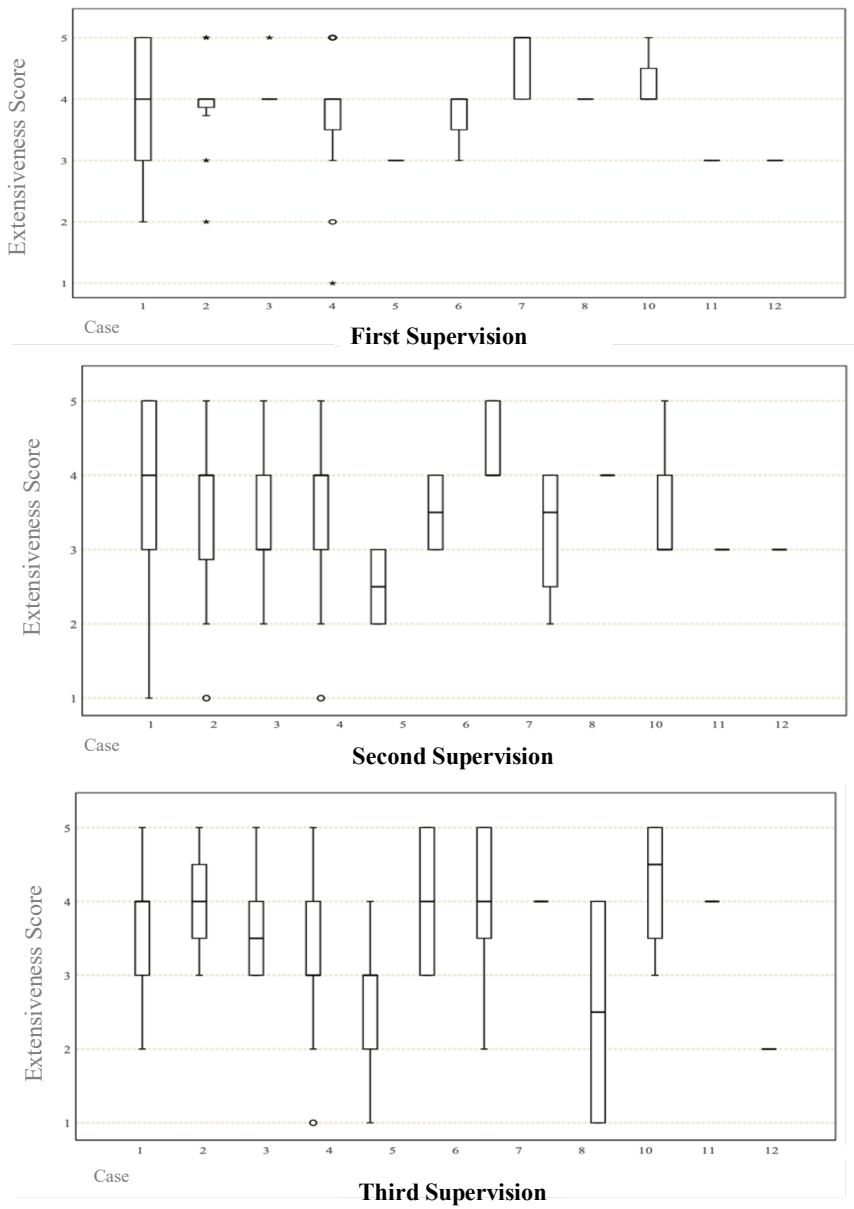


Figure 3

Efficiency Across Cases



References

Accurso, E. C., Taylor, R. M., & Garland, A. F. (2011). Evidence-based practices addressed in community-based children's mental health clinical supervision. *Training and Education in Professional Psychology, 5*(2), 88.

Altman, D. G. (1990). *Practical statistics for medical research.* Chapman and Hall/CRC.

Andrews, G., Basu, A., Cuijpers, P., Craske, M., McEvoy, P., English, C., & Newby, J. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. *Journal of anxiety disorders, 55*, 70-78.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics, 34*(4), 555-596.

Bailin, A., Bearman, S. K., & 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*(5), 800-812.

Barican, J. L., Yung, D., Schwartz, C., Zheng, Y., Georgiades, K., & Waddell, C. (2022). Prevalence of childhood mental disorders in high-income countries: A systematic review and meta-analysis to inform policymaking. *Evidence-Based Mental Health, 25*(1), 36-44.

Bearman, S. K., Weisz, J. R., Chorpita, B. F., Hoagwood, K., Ward, A., Ugueto, A. M., & Bernstein, A. (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*(6), 518-529.

Becker, K. D., Boustani, M., Gellatly, R., & Chorpita, B. F. (2018). Forty years of engagement research in children's mental health services: Multidimensional measurement and practice elements. *Journal of Clinical Child & Adolescent Psychology, 47*(1), 1-23.

Becker-Haimes, E. M., Lushin, V., Creed, T. A., & Beidas, R. S. (2019). Characterizing the heterogeneity of clinician practice use in community mental health using latent profile analysis. *BMC psychiatry, 19*(1), 1-11.

Becker-Haimes, E. M., Okamura, K. H., Baldwin, C. D., Wahesh, E., Schmidt, C., & Beidas, R. S. (2019). Understanding the landscape of behavioral health pre-service training to inform evidence-based intervention implementation. *Psychiatric services*, *70*(1), 68–70.

Beidas, R. S., Edmunds, J. M., Marcus, S. C., & Kendall, P. C. (2012). Training and consultation to promote implementation of an empirically supported treatment: A randomized trial. *Psychiatric Services, 63*(7), 660-665.

Beidas, R. S., & Kendall, P. C. (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.

Beidas, R. S., Williams, N. J., Becker-Haimes, E. M., Aarons, G. A., Barg, F. K., Evans, A. C., Jackson, K., Jones, D., Hadley, T., & Hoagwood, K. (2019). A repeated cross-sectional study of clinicians' use of psychotherapy techniques during 5 years of a system-wide effort to implement evidence-based practices in Philadelphia. *Implementation Science, 14*(1), 1-13.

Bernstein, A., Chorpita, B. F., Daleiden, E. L., Ebesutani, C. K., & Rosenblatt, A. (2015). Building an evidence-informed service array: Considering evidence-based programs as well as their practice elements. *Journal of Consulting and Clinical Psychology, 83*(6), 1085.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied research in Memory and Cognition, 9*(4), 475.

Borghouts, J., Eikey, E., Mark, G., De Leon, C., Schueller, S. M., Schneider, M., Stadnick, N., Zheng, K., Mukamel, D., & Sorkin, D. H. (2021). Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. *Journal of medical Internet research, 23*(3), e24387.

Brabson, L. A., Harris, J. L., Lindhiem, O., & Herschell, A. D. (2020). Workforce turnover in community behavioral health agencies in the USA: A systematic review with recommendations. *Clinical child and family psychology review, 23*(3), 297-315.

Bradley, W. J., & Becker, K. D. (2021). Clinical supervision of mental health services: A systematic review of supervision characteristics and practices associated with formative and restorative outcomes. *The Clinical Supervisor, 40*(1), 88-111.

Brookman-Frazee, L., Stadnick, N. A., Lind, T., Roesch, S., Terrones, L., Barnett, M. L., Regan, J., Kennedy, C. A., F Garland, A., & Lau, A. S. (2021). Therapist-observer concordance in ratings of EBP strategy delivery: Challenges and targeted directions in pursuing pragmatic measurement in children's mental health services. *Administration and Policy in Mental Health and Mental Health Services Research, 48*(1), 155-170.

Brookman-Frazee, L. I., Drahota, A., & Stadnick, N. (2012). Training community mental health therapists to deliver a package of evidence-based practice strategies for school-age children with autism spectrum disorders: A pilot study. *Journal of autism and developmental disorders, 42*(8), 1651-1661.

Chorpita, B., Becker, K., & Park, A. (2018). Action cycle and use of evidence behavioral observation coding system (ACE-BOCS)[Measurement instrument]. Unpublished instrument.

Chorpita, B. F., & Daleiden, E. L. (2014). Structuring the collaboration of science and service in pursuit of a shared vision. *J Clin Child Adolesc Psychol, 43*(2), 323-338.

Chorpita, B. F., Daleiden, E. L., Malik, K., Gellatly, R., Boustani, M. M., Michelson, D., Knudsen, K., Mathur, S., & Patel, V. H. (2020). Design process and protocol description for a multi-problem mental health intervention within a stepped care approach for adolescents in India. *Behaviour Research and Therapy, 133*, 103698.

Chorpita, B. F., Daleiden, E. L., Vera, J. D., & Guan, K. (2021). Creating a prepared mental health workforce: comparative illustrations of implementation strategies. *Evidence-Based Mental Health, 24*(1), 5-10.

Cicchetti, D. V. (2001). Methodological commentary the precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23*(5), 695-700.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128–152.

Collatz, L., & Wetterling, W. (2012). *Optimization problems.* Springer Science & Business Media.

Deming, W. E. (1989). *Out of the Crisis. Quality, productivity and competitive position.* Massachusetts Institute of Technology, Cambridge, MA, 81, 82.

Dorsey, S., Pullmann, M. D., Kerns, S. E., 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*(6), 838-852.

Edmunds, J. M., Kendall, P. C., Ringle, V. A., Read, K. L., Brodman, D. M., Pimentel, S. S., & Beidas, R. S. (2013). An examination of behavioral rehearsal during consultation as a predictor of training outcomes. *Administration and Policy in Mental Health and Mental Health Services Research, 40*(6), 456-466.

Falender, C. A., & Shafranske, E. P. (2004). *Clinical supervision: A competency-based approach.*

Frank, H. E., Becker-Haimes, E. M., & Kendall, P. C. (2020). Therapist training in evidence-based interventions for mental health: a systematic review of training approaches and outcomes. *Clinical Psychology: Science and Practice, 27*(3), e12330.

Fukui, S., Rapp, C. A., Goscha, R., Marty, D., & Ezell, M. (2014). The perceptions of supervisory support scale. *Administration and Policy in Mental Health and Mental Health Services Research, 41*(3), 353-359.

Garland, A. F., Bickman, L., & Chorpita, B. F. (2010). Change what? Identifying quality improvement targets by investigating usual mental health care. *Administration and Policy in Mental Health and Mental Health Services Research, 37*(1), 15-26.

Garland, A. F., Hurlburt, M. S., Brookman-Frazee, L., Taylor, R. M., & Accurso, E. C. (2010). Methodological challenges of characterizing usual care psychotherapeutic practice. *Administration and Policy in Mental Health and Mental Health Services Research, 37*(3), 208-220.

Glasgow, G. E. (1999). *Issue publics in American politics California Institute of Technology*.

Graham, I. D., Logan, J., Harrison, M. B., Straus, S. E., Tetroe, J., Caswell, W., & Robinson, N. (2006). Lost in knowledge translation: time for a map? *Journal of continuing education in the health professions, 26*(1), 13-24.

Herschell, A. D., Kolko, D. J., Baumann, B. L., & Davis, A. C. (2010). The role of therapist training in the implementation of psychosocial treatments: A review and critique with recommendations. *Clinical psychology review, 30*(4), 448-466.

Heyman, R. E., Otto, A. K., Reblin, M., Wojda, A. K., & Xu, S. (2021). The lump-versus-split dilemma in couple observational coding: A multisite analysis of rapid marital interaction coding system data. *Journal of Family Psychology, 35*(4), 559.

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research, Practice, Training, 33*(2), 332.

Hunsley, J., & Mash, E. J. (2020). The role of assessment in evidence-based practice. *Handbook of assessment and treatment planning for psychological disorders*, 3-22. The Guilford Press.

Kaltenthaler, E., Sutcliffe, P., Parry, G., Beverley, C., Rees, A., & Ferriter, M. (2008). The acceptability to patients of computerized cognitive behaviour therapy for depression: a systematic review. *Psychological medicine, 38*(11), 1521-1530.

Kazdin, A. E. (2019). Annual Research Review: Expanding mental health services through novel models of intervention delivery. *Journal of Child Psychology and Psychiatry, 60*(4), 455-472.

Knudsen, K. S., Becker, K. D., Guan, K., Gellatly, R., Patel, V. H., Malik, K., Boustani, M. M., Mathur, S., & Chorpita, B. F. (2021). A pilot study to evaluate feasibility and acceptability of training mental health workers in India to select case-specific intervention procedures within a dynamic modular treatment designed for a low-resource setting. *Journal of Evaluation in Clinical Practice, 28*(4), 531-541.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155-163.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 159*(174).

Lucid, L., Meza, R., Pullmann, M. D., Jungbluth, N., Deblinger, E., & Dorsey, S. (2018). Supervision in community mental health: Understanding intensity of EBT focus. *Behavior Therapy, 49*(4), 481-493.

Lyon, A. R., Dorsey, S., Pullmann, M., Silbaugh-Cowdin, J., & Berliner, L. (2015). Clinician use of standardized assessments following a common elements psychotherapy training and consultation program. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(1), 47-60.

Lyon, A. R., Pullmann, M. D., Whitaker, K., Ludwig, K., Wasse, J. K., McCauley, E., & Dorsey, S. (2021). Standards of evidence for digital mental health interventions: Empirical examination of the Institute of Medicine's evaluative framework. *Depression and anxiety, 38*(7), 754-763.

Lyon, A. R., Pullmann, M. D., & Stadnick, N. A. (2019). Community mental health services for children and adolescents. *Child and Adolescent Psychiatry Clinics, 28*(1), 13-23.

McHugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *American Psychologist, 65*(2), 73.

Mellado, D. H., & Suarez, S. L. (2008). The connection between the training supervisor and the clinical supervisor in psychotherapy. *Psychodynamic Practice, 14*(4), 419-433.

National Research Council. (2014). *Improving the utility and translation of animal models for nervous system disorders: Workshop summary*. National Academies Press.

Nelson, M. M., & Steele, R. G. (2007). Predictors of practitioner self-reported use of evidence-based practices: Practitioner training, clinical setting, and attitudes toward research. *Administration and Policy in Mental Health and Mental Health Services Research, 34*(4), 319-330.

O'Connor, E., Senger, C. A., Henninger, M. L., Coppola, E., & Gaynes, B. N. (2019). Interventions to prevent perinatal depression: Evidence report and systematic review for the US Preventive Services Task Force. *Jama, 321*(6), 588-601.

Orlinsky, D. E., Ronnestad, M. H., & Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*, 307-390.

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of Global Health, 8*(2).

Parikh, S. V., Huniewicz, P., E-health: an overview of the uses of the Internet, social media, apps, and websites for mood disorders. *Current Opinion in Psychiatry, 32*(5), 413-417.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(2), 65-76.

Raghavan, R., Bright, C. L., & Shadoin, A. L. (2008). Toward a policy ecology of implementation of evidence-based practices in public mental health settings. *Implementation Science, 3*(1), 1-14. Reinholt, N., & Krogh, L. (2014). Efficacy of transdiagnostic cognitive behaviour therapy for anxiety disorders: A systematic review and meta-analysis of published outcome studies. *Cognitive Behaviour Therapy, 43*(3), 171-184.

Roane, D. M., Lerman, D. C., Kelley, M. E., & Van Camp, C. M. (1999). Within-session patterns of

compliance in the compliance training of young children. *Journal of Applied Behavior Analysis, 32*(1), 65-76.

Sachs-Ericsson, N., Sheffler, J., Stanley, I. H., Piazza, J. R., & Preacher, K. J. (2017). When emotional pain becomes physical: Adverse childhood experiences, pain, and the role of mood and anxiety disorders. *Journal of Clinical Psychology, 73*(10), 1403-1428.

Schoenwald, S. K., & Henggeler, S. W. (2004). A public health perspective on the transport of evidence-based practices. *Clinical psychology: Science and practice, 11*(4), 360-363.

Schoenwald, S. K., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychiatric Services, 52*(9), 1190-1197.

Schoenwald, S. K., Sheidow, A. J., & Chapman, J. E. (2009). Clinical supervision in treatment transport: effects on adherence and outcomes. *Journal of consulting and clinical psychology, 77*(3), 410.

Seligman, M. E., Rashid, T., & Parks, A. C. (2006). Positive psychotherapy. *American Psychologist, 61*(8), 774.

Shapiro, J., Brown, M., & Dean, K. (2019). An exploratory study examining supervision and ethical dilemmas. *Clinical Social Work Journal, 47*(3), 232-240.

Shiner, B., Leonard Westgate, C., Bernardy, N. C., Thompson, P., & Watts, B. V. (2020). Trends in prevalence of depression and antidepressant prescribing in children and adolescents: A cohort study in the Health Improvement Network (THIN). *PLoS One, 15*(1), e0228016.

Skinner, C. H., & Beatty, J. (2016). Technology-based intervention and instruction delivery in rural schools: A review of research. *Contemporary School Psychology, 20*(3), 221-233.

Stadnick, N. A., Roesch, S. C., & Barnett, M. L. (2020). Communicating fidelity in strategic family therapy for adolescent behavior problems: Measurement application and comparison across therapy phases. *Journal of Marital and Family Therapy, 46*(2), 277-290.

Stirman, S. W., Gutner, C. A., Langdon, K., & Graham, J. R. (2015). Bridging the gap between research and practice in mental health service settings: An overview of developments in implementation theory and research. *Behavior Therapy, 47*(6), 920-936.

Stirman, S. W., Miller, C. J., Toder, K., & Calloway, A. (2013). Development of a framework and coding system for modifications and adaptations of evidence-based interventions. *Implementation Science, 8*(1), 1-9.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist, 62*(4), 271.

Suveg, C., Hudson, J. L., Brewer, G. A., Flannery-Schroeder, E., & Gosch, E. A. (2009). Cognitive-behavioral therapy for anxiety-disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of Consulting and Clinical Psychology, 77*(2), 282.

Valenstein-Mah, H., & Dorsey, S. (2010). The leadership training program in mental health and addiction services: an innovative approach to increasing supervisor knowledge and skill. *Journal of Mental Health, 19*(3), 255-263.

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., ... & Wells, J. E. (2007). Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *The Lancet, 370*(9590), 841-850.

Wasserman, G. A., McReynolds, L. S., Ko, S. J., Katz, L. M., Carpenter, J. R., & Cauffman, E. (2005). Gender differences in psychiatric disorders at juvenile probation intake. *American Journal of Public Health, 95*(1), 131-137.

Weisz, J. R., Ng, M. Y., Bearman, S. K., Odd couple? Reenvisioning the relation between science and practice in the dissemination-implementation era. *Clinical Psychological Science, 3*(1), 58-74.

Wolfe, D. A., & Crooks, C. V. (2001). A comprehensive strategy to stop violence against women: Canada's transition from national policy to local practice. *Trauma, Violence, & Abuse, 2*(1), 76-85.

Zima, B. T., Hurlburt, M. S., Knapp, P., Ladd, H., Tang, L., Duan, N., Wallace, P., & Rosenblatt, A. (2005). Quality of publicly-funded outpatient specialty mental health care for common childhood psychiatric disorders in California. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(2), 130-144.

Zimmerman, F. J., Christakis, D. A., & Meltzoff, A. N. (2007). Television and DVD/video viewing in children younger than 2 years. *Archives of Pediatrics & Adolescent Medicine, 161*(5), 473-479.

Zozaya, L. E., Figley, C. R., & Mcgarity, S. (2017). A study of burnout and turnover issues and prevention strategies. *Journal of Health & Human Services Administration, 39*(3), 298-333.