Spring 2023

Assessing the Reliability of the Five-in-20 Classroom Observation Tool

Danielle Buechlein
Eastern Illinois University

Assessing the Reliability of the Five-in-20 Classroom Observation Tool

Abstract

Effective classroom management practices are crucial to fostering a positive learning environment, student achievement, and student social-emotional development, as well as teacher job satisfaction. The Five-in-20 Classroom Observation Tool (FCOT) was developed from the 21 evidence-based strategies identified by Simonsen and colleagues (2008) to assess and support teachers' classroom management; however, its psychometric properties are unknown. The present study looks at the interobserver agreement of the FCOT, specifically at how consistent observers' ratings are with each other and how consistent observers' ratings are with a standard coded copy of the FCOT. Forty-four participants, using the FCOT, rated the same 20-minute video of classroom instruction. Their ratings were compared using intraclass correlation coefficient and alpha coefficient. Intraclass correlation coefficient yielded excellent agreement for average measures and moderate agreement for single measures for both strategy endorsement and quality rating. Cronbach's alpha also suggested strong reliability for strategy and quality. The participants' ratings were compared to the standard coded FCOT using Cohen's kappa and percent agreement. The kappa value for strategy endorsement suggested substantial agreement, while the kappa value for quality rating suggested moderate agreement. Percent agreement for strategy endorsement was calculated at 85.37% and for quality rating at 60.24%. Implications and suggestions for future research are provided.

Degree Type

Dissertation/Thesis

Degree Name

Specialist in School Psychology

Department

Psychology

Thesis Director

Margaret T. Floress

Thesis Committee Member

Gary L. Canivez

Thesis Committee Member

Hao-Jan Luh

Keywords

classroom management, direct observation, evidence-based strategies, consultation, reliability

Subject Categories

Curriculum and Instruction | Early Childhood Education | Elementary Education | Secondary Education | Special Education and Teaching | Teacher Education and Professional Development

Assessing the Reliability of the Five-in-20 Classroom Observation Tool

Danielle Buechlein

Eastern Illinois University

Abstract

Effective classroom management practices are crucial to fostering a positive learning environment, student achievement, and student social-emotional development, as well as teacher job satisfaction. The Five-in-20 Classroom Observation Tool (FCOT) was developed from the 21 evidence-based strategies identified by Simonsen and colleagues (2008) to assess and support teachers' classroom management; however, its psychometric properties are unknown. The present study looks at the interobserver agreement of the FCOT, specifically at how consistent observers' ratings are with each other and how consistent observers' ratings are with a standard coded copy of the FCOT. Forty-four participants, using the FCOT, rated the same 20-minute video of classroom instruction. Their ratings were compared using intraclass correlation coefficient and alpha coefficient. Intraclass correlation coefficient yielded excellent agreement for average measures and moderate agreement for single measures for both strategy endorsement and quality rating. Cronbach's alpha also suggested strong reliability for strategy and quality. The participants' ratings were compared to the standard coded FCOT using Cohen's kappa and percent agreement. The kappa value for strategy endorsement suggested substantial agreement, while the kappa value for quality rating suggested moderate agreement. Percent agreement for strategy endorsement was calculated at 85.37% and for quality rating at 60.24%. Implications and suggestions for future research are provided.

Keywords: classroom management, direct observation, evidence-based strategies, consultation, reliability

Acknowledgements

I want to thank Dr. Floress for her consistent guidance and helpful advice throughout the thesis process. I am appreciative of every edit and suggestion that contributed to the final product. Thank you for providing brainstorming sessions and nudging me in the right direction. I also want to thank my committee members, Dr. Canivez and Dr. Luh, whose edits and suggestions during my proposal were the spark I needed for my analysis. Thank you to those who took time to participate, especially to those who provided feedback related to the tool. Finally, thank you to my fiancé for supporting me through graduate school and for constantly cheering me on.

Table of Contents

Tables

Assessing the Reliability of the Five in 20 Classroom Observation Tool

Direct observations are empirically sound and commonly used for assessing behavioral, social, and emotional problems in school-aged children and adolescents. Direct observations are also used to support teacher professional development and training (Wilson & Reschly, 1996). This is particularly important because effective classroom management is critical for student and teacher success, and teachers receive limited preservice and in-service support in classroom management (Freeman et al., 2014; Oliver & Reschly, 2010). Direct observations are carried out by measuring specific behaviors using operational definitions developed a priori. Procedures and scoring are standardized and do not vary from one observer to another (Hintze & Matthews, 2004).

Researchers have commonly used accuracy and interobserver agreement (IOA) when addressing issues of reliability related to direct observation (Hintze & Matthews, 2004). However, few (if any) available classroom management observation tools (MacSuga & Simonsen, 2011; Simonsen et al., 2008; Missouri School-Wide Positive Behavior Support [MO SWPBIS], 2017) have been carefully studied to document psychometric properties. In other words, the reliability and validity of many tools available to consultants have not been established. Other tools, which have established psychometric properties, such as the *Classroom Assessment and Scoring System* (CLASS; Pianta et al., 2008), are resource-intensive for school-based consultants to use without additional training and support. Therefore, there is a need for simple, brief observation tools that efficiently allow an observer (e.g., school psychologist, behavioral consultant, instructional coach, administrator) to assess a teacher's use of evidence-based classroom management strategies, with adequate reliability and validity. The following section will discuss key aspects of classroom management.

Classroom Management

Classroom management and the degree to which it is implemented properly impacts many aspects in the school and is thus a crucial part of teaching effectively (Aloe et al., 2014; Freeman et al., 2014; Simonsen et al., 2008). Classroom management occurs before children enter the classroom (Simonsen et al., 2008) and encompasses teacher actions within and outside of direct instruction and facilitates academic and social-emotional learning (Garwood et al., 2017). Despite its importance, definitions of effective classroom management as derived from research differ greatly from those used in teacher training. Many states in the U.S. require preservice teachers to receive training in classroom management, however the required training often does not include evidence-based classroom management practices (Freeman et al., 2014). Furthermore, because student-teachers typically start student teaching after the school year is underway, many novice teachers miss the crucial opportunity to build their own classroom management systems or to develop management strategies (Garwood et al., 2017). The next section discusses why classroom management is critical to effective teacher instruction.

Why is Classroom Management Important?

Classroom management is critical to teaching because it positively impacts the learning environment, student achievement, and student social-emotional development (Lekwa et al., 2018). Student behavior and teacher practices are included as two crucial components of the learning environment (Lekwa et al., 2018). Effective teachers use critical classroom management skills (e.g., behavior management, instruction, and addressing student concerns), which are tied to her or his ability to teach effectively (Freeman et al., 2014). In well-run classrooms, teachers are likely to feel accomplished as their students learn, whereas excessive student misbehavior can lead to a chaotic classroom environment and low teacher satisfaction and accomplishment.

These and other similar feelings, such as emotional exhaustion and depersonalization, have previously been connected to teacher burnout (Aloe et al., 2014).

Positive Outcomes for Students and Teachers

Student Achievement. Effectively managed classrooms have positive effects on student achievement. For instance, teachers who effectively manage student classroom behavior have more time to teach, because they are spending less time addressing student misbehavior (Garwood et al., 2017). Fruth (2014) examined the impact of the PAX Good Behavior Game (a classroom management system) 4th grade students' behavior from one Midwestern school. Twenty-six students (one classroom) received intervention while 57 students (two classrooms) were in the control condition. The authors concluded the increased structure and predictable learning environment (provided by the PAX Good Behavior Game) allowed for more effective instruction and positively influenced students' learning compared to those in the control classrooms.

Social-Emotional Development. In addition to impacting student achievement, effective universal classroom management supports the development of students' appropriate social behavior. Effective classroom management strategies positively impact students' social-emotional development (Reinke et al., 2021). Less effective practices are correlated with increased levels of behavioral disruption (Stronge et al., 2011) and decreased levels of student on-task behavior (Aloe et al., 2014). Furthermore, programs focusing on social-emotional behaviors have been shown to positively impact aggressive behaviors, academic performance, and substance use (Domitrovich et al., 2016).

Without effective classroom management, teachers are often hindered in addressing problematic behaviors, and may resort to less effective, punitive, and inconsistent management

strategies (Garwood et al., 2017). For these reasons, effective classroom management practices may be especially important for students at risk for emotional and behavioral disorders. For example, Garwood and colleagues (2017) examined the effects of evidence-based classroom management (measured using the Classroom Assessment Scoring System) on reading achievement of students in kindergarten through third grade with or at risk for emotional and behavioral disorders. The researchers found male students' reading achievement (as measured with the Passage Comprehension and Letter-Word Identification subtests from the Woodcock Johnson III) was significantly related to effective classroom management practices. With the passage of the Individuals with Disabilities Education Improvement Act (IDEIA, 2004) students with behavioral challenges and other disabilities receive instruction in the general education classroom more often than in the past (Freeman et al., 2014). Therefore, all students in the general education classroom benefit when teachers are equipped with effective classroom management strategies.

Students with Learning and Behavioral Needs. Children identified as having learning, emotional, and behavioral disabilities alike especially benefit from effective classroom management. Children with emotional or behavioral disorders often display disruptive behaviors, internalizing and/or externalizing in nature, in the classroom environment (Garwood et al., 2017), suggesting the importance of addressing these behaviors to maximize learning potential. Aggressive behaviors, for example, pose a threat to student outcomes if not addressed. Without successful intervention, childhood aggression can lead to a myriad of adverse outcomes, most immediately impeding learning; however, improving teacher classroom management has shown to improve math achievement in students who are aggressive (Chuang et al., 2020).

Children with disabilities further benefit from specific classroom management strategies, such as those that support appropriate social behaviors including communication skills, problem-solving abilities, and coping skills (Reinke et al., 2021). The Incredible Years Teacher Classroom Management program (IY TCM) is an example of a universal social-emotional intervention of particular interest for school professionals working with children with disabilities. Reinke and colleagues (2021) found support for their hypothesis that special education children placed in IY TCM classrooms would show improvements in social behaviors, namely disruptive behavior and social competence, compared to control participants.

Teacher Burnout. Teachers also benefit from implementing effective classroom management strategies. Given student misbehavior contributes to teacher burnout, ensuring teachers utilize effective management practices may help prevent teachers from leaving the field (Ingersoll, 2001). Ingersoll (2001) extracted data from the National Center for Education Statistics Schools and Staffing Survey and the Teacher Follow up Survey, which includes answers from administrators and teachers. Twenty-five percent of the teachers who left the field reported they left due to job dissatisfaction. Of these teachers, 30% reported leaving due to student discipline problems (Ingersoll, 2001). Many teachers report feeling unprepared to effectively manage student behavior and mental health concerns (Reinke et al., 2011). Feeling unprepared can negatively impact teachers' sense of self-efficacy, which is a protective factor against burnout (Aloe et al., 2014).

Aloe and colleagues (2014) conducted a meta-analysis examining in-service teachers and burnout. Sixteen articles were included in the analysis and the authors found a moderate relation between classroom management self-efficacy and dimensions of burnout, including increased emotional exhaustion, elevated feelings of depersonalization, and decreased feelings of personal

accomplishment. The impact of teacher burnout also negatively influences students. For example, work ability, mental health, and emotion regulation have all been linked to burnout, and teachers experiencing these feelings are more likely to foster negative social environments for their students (Aloe et al., 2014). Teachers need to know how to create healthy classroom environments (i.e., in part by implementing effective classroom management; Domitrovich et al., 2016).

Domitrovich and colleagues (2016) demonstrated how improving student behavior positively influenced teacher self-efficacy and burnout. Specifically, teachers were trained to implement the PAX Good Behavior Game (a classroom behavior intervention) and Promoting Alternative Thinking Strategies (PATHS; a social-emotional curriculum). Results demonstrated that when teachers implemented these social-emotional practices, social-emotional competence improved (i.e., higher teacher efficacy in both behavior and SEL management; Domitrovich et al., 2016). Positive outcomes for both students and teachers emphasize the importance of training and supporting teachers to use effective classroom management practices.

Critical Features of Classroom Management

Lekwa and colleagues (2018) define behavioral classroom management strategies as strategies teachers use to increase student compliance with tasks and expectations, either proactively or after the behavior occurs. Sugai and Horner (2002) defined effective classroom management as strategies that maximize instruction time, maximized academic engagement and achievement, and proactive management strategies. Other effective strategies highlighted in the research literature include praise, clear expectations, routines and schedules, consistent consequences, and ignoring (Reinke et al., 2021). The present study drew from the work of

Simonsen and colleagues (2008) surrounding 21 general, evidence-based practices that fall into five evidence-based critical features of classroom management, which are outlined next.

Maximize Structure. The first of the five critical features refer to the use of teacher-directed activity, explicitly defined routines, and the physical structure of the classroom (Simonsen et al., 2008). The physical structure of the classroom includes permanent features that define the space, the placement of furniture that impacts seating arrangements and traffic flow, and visual stimuli displayed on the walls. Classrooms with higher-quality structure have been linked to increases in appropriate academic and social behaviors, such as on-task behavior, positive peer interactions, helpful behaviors, and decreased aggression (Simonsen et al., 2008).

Another aspect of maximizing structure in the classroom is minimizing crowding and distractions. Students with more space, regardless of overall room design, increased positive interactions with peers, teachers, and parents (Simonsen et al., 2008). More walls or visual dividers in a classroom has been shown to decrease teacher and student distraction and increase student satisfaction and classroom activities. Altering classrooms has shown efficacy in increasing the variety of appropriate behaviors (Simonsen et al., 2008).

Post, Teach, Review, Monitor, and Reinforce Expectations. When setting expectations, teachers should identify and define a reasonable number of rules, stated positively and broadly enough to include all desired behaviors (Simonsen et al., 2008). These rules, or expectations, should be clearly posted in the classroom and explicitly taught to students. Additionally, the teacher should frequently review the classroom expectations with students and actively monitor and supervise for adherence. Posting, teaching, reviewing, monitoring, and reinforcing expectations has been associated with lower rates of off-task and disruptive behavior and an increase in academic engagement, leadership behaviors, and conflict resolution. Further,

research shows that pairing the instruction of the rules with feedback and reinforcement produces the largest gains (Simonsen et al., 2008).

The aspect of active supervision included within this critical feature of classroom management is also heavily supported in the literature. Active supervision has been demonstrated to have positive impacts on student behaviors, such as higher participation and decreased numbers of minor incidents, across settings. Research has also shown that it is the degree to which the teacher employs active supervision, rather than the teacher-to-student ratio, that accounts for the most variance in problem behaviors exhibited in non-classroom settings (Simonsen et al., 2008).

Actively Engage Students in Observable Ways. The term engagement refers to how a student participates in classroom instruction (both actively and passively). In the literature, engagement is the strongest mediator between instruction and achievement (Simonsen et al., 2008). Teachers can increase active engagement by employing five general practices, including increasing opportunities to respond, using direct instruction techniques, implementing peer tutoring, utilizing computer-based instruction, and providing guided notes. Increasing opportunities to respond, or teacher behavior meant to solicit a response from students including choral responding and response cards, has positive effects on student achievement and behavior (i.e., on-task behavior, engagement, correct responses; Simonsen et al., 2008).

The second general practice included within this critical feature, direct instruction, encompasses a clear, sequenced, and supported presentation of content and instruction, as well as increased rates of opportunities to respond, review of content and associated feedback, and progress monitoring (Simonsen et al., 2008). Direct instruction also has ample support in the research literature. Impressive gains have been reported on measures of basic skill, self-esteem,

and cognitive reasoning. Direct instruction has also contributed to gains in academic achievement and higher rates of on-task behaviors (Simonsen et al., 2008).

The third general practice included within this critical feature is implementing peer tutoring. To utilize peer tutoring, teachers pair students and assign one to the role of tutor and the other to the role of tutee. The teacher then moves around the classroom to assist individual pairs as needed. Class-wide peer tutoring increases engagement, reading achievement, and on-task behaviors (Simonsen et al., 2008). The fourth general practice is computer-assisted instruction, which provides one-on-one instruction to every student in the classroom. Computer-assisted instruction is particularly effective for students with ADHD. In this population, computer-assisted instruction has increased active engagement in math instruction, oral reading fluency in reading instruction, and on-task behaviors across areas (Simonsen et al., 2008).

The final general practice within this critical feature is providing guided notes to students. Guided notes are outlines of a given lecture or book chapter provided by the teacher. The outlines contain main ideas of the material with spaces open for students to fill in missing information. Using guided notes is based on the notion that students who take frequent and relevant notes during instruction learn more than passive students. This practice is associated with increases in academic achievement (Simonsen et al., 2008).

Use of a Continuum of Strategies to Acknowledge Appropriate Behavior. Teachers can use a variety of empirically supported strategies tailored to the identification and recognition of appropriate classroom behaviors, including specific praise, group reinforcement contingencies, behavior contracts, and token economies (Simonsen et al., 2008). Each of these strategies are supported by research evidence, however praise has the strongest evidence base. Delivering contingent praise in response to academic behavior has increased participants'

productivity, accuracy, correct responses, and academic performance. Likewise, contingent praise for appropriate social behavior increases on-task behavior, student attention and compliance, positive self-statements, and cooperative play (Simonsen et al., 2008).

Group reinforcement contingencies and token economies are also adequately supported in the literature. These strategies have demonstrated increased positive verbal interactions, appropriate classroom behavior, peer acceptance, achievement, and student attention. Additionally, group reinforcement contingencies and token economies decrease negative verbal interactions, transition times, inappropriate behaviors, out-of-seat-behavior, and talk-outs (Simonsen et al., 2008). Behavior contracts also increase student productivity, on-task behavior, and task completion, as well as improved grades and self-control (Simonsen et al., 2008).

Use a Continuum of Strategies to Respond to Inappropriate Behavior. Inappropriate behaviors in the classroom can be addressed using a range of simple to complex evidence-based strategies intended to decrease the likelihood of these behaviors in the future. The strategies indicated by Simonsen and colleagues (2008) are brief, contingent, and specific error corrections; performance feedback; differential reinforcement; planned ignoring; response cost; and time out from reinforcement. Brief, contingent, and specific error corrections are informative statements provided by the teacher when a student engages in an undesired behavior. The teacher briefly and concisely states the behavior and instructs the student on what she or he should do in the future. The literature has demonstrated that quiet, brief, and consistent corrections are the most effective (Simonsen et al., 2008).

Performance feedback is when students are given feedback in response to engagement in a target behavior (Simonsen et al., 2008). Teachers present data to students in a visual manner to assist in analyzing changes in performance. Performance feedback (particularly public posting)

surrounding target social behaviors increases appropriate classroom behaviors and decreased the frequency of targeted behaviors, decreased transition times, and increased positive social and academic behaviors (Simonsen et al., 2008).

Differential reinforcement is implemented by providing contingent reinforcement to the student when he/she engages in appropriate behavior and not attending to undesired behaviors. Differential reinforcement increases appropriate behaviors while decreasing inappropriate behaviors (Simonsen et al., 2008). Differential reinforcement is often used in combination with planned ignoring, which is when the teacher withholds attention from a student engaging in an undesired behavior. Planned ignoring has demonstrated positive effects on social behavior and study habits (Simonsen et al., 2008).

The final strategies used to address inappropriate behaviors are response cost and time out from reinforcement. Response cost is when a desired stimulus is removed when the student engages in an inappropriate behavior, which has effectively decreased swearing, aggression, and other inappropriate behaviors (Simonsen et al., 2008). Time out from reinforcement is implemented by removing the student from reinforcement when they engage in an undesired behavior and has a strong evidence-based in the literature (Simonsen et al., 2008). Despite the research support for implementing evidence-based, behavior management strategies; teacher preparation programs often do not incorporate this content into the curriculum (Freeman et al., 2014; Garwood et al., 2017). Furthermore, implementing behavior management strategies post-training is highly variable and often discrepant from empirically based recommendations (Lekwa et al., 2018). School psychologists can provide behavior management consultation to train and support teachers in using evidence-based strategies effectively and more consistently.

Consultation and Importance of Supporting Teachers

Consultation in schools is a process where a professional provides psychological and educational services, while cooperatively working with a staff member to improve a student's, or several students' learning (Erchul & Martens, 2010). Consultation often takes the form of face-to-face interactions during which the consultant (e.g., school psychologist) guides the consultee (e.g., teacher) utilizing systematic problem-solving, professional support, and social influence. The consultee then uses the tools gained during consultation to help the client (e.g., student), via intervention (Erchul & Martens, 2010). The first steps in the problem-solving process are problem identification and analysis (Erchul & Martens, 2010) and direct observation is used to collect these data. Additionally, direct observation is used to determine whether an intervention was implemented with fidelity and whether it was effective (Freeman et al., 2014).

School psychologists are trained to provide consultation, which is a crucial part of the response to intervention model (Kratochwill et al., 2014). As consultants, school psychologists address teacher and parent concerns, as well as system-wide issues. In a survey of school psychologists, participants rated themselves as highly knowledgeable in consultation, likely resulting from the National Association of School Psychologists' (NASP) push for increasing consultative services (Bahr et al., 2017). During consultation, the school psychologist builds relationships with consultees, identifies and defines problems, establishes goals, analyzes problems, and plans interventions. School psychologists also support consultees by building consultees' skills, monitoring intervention implementation, gathering progress monitoring data, and revising the intervention plan if necessary (Kratochwill et al., 2014). By supporting consultees (i.e., teachers) via consultation, this process has the potential to help retain high quality teachers.

How Does Consultation Help Retain Quality Teachers?

Retaining quality teachers has become an increasingly important task for schools, however it is an issue that has received little attention in the past (Albright et al., 2017). Compared to previous generations, American teachers today are subjected to greater levels of stress, partially due to increasing diversity (i.e., SES, nationality, language, etc.) and decreasing parental involvement (Aloe et al., 2014). This stress, often compounded by feelings of anxiety and inadequacy from a lack of training and low levels of support, impacts teachers' health, job satisfaction, and ultimately contributes to the country's elevated teacher attrition rate (Prilleltensky et al., 2016). Teachers may also develop feelings of exhaustion, uncertainty, and burnout due to constantly evolving state and federal standards and a lack of support within schools (Aloe et al., 2014). As a result, the United States is experiencing an epidemic of teachers leaving the field.

Rates of teacher attrition vary from source to source and across regions, however the literature estimates that anywhere from 17% to 50% of teachers leave the field within the first five years (Albright et al., 2017; Domitrovich et al., 2016; Freeman et al., 2014; Prilleltensky et al., 2016). In addition to the factors stated previously, teachers consistently leave the field for similar reasons (e.g., lack of training, poor school environment, and poor student behavior; Freeman et al., 2014). Poor student behavior and struggling with classroom management seem to be common reasons teachers leave the field (Prilleltensky et al., 2016). Teachers also report leaving due to little upper-level administrative support, a lack of adequate resources, government policies, and attitudes in the community (Albright et al., 2017).

The issue of teachers having a lack of training is not due to a lack of effort, but rather a discrepancy in what is taught in teacher preparation programs and what is practiced in schools.

Responses from two focus groups of new and experienced teachers indicated that coursework related to classroom management did not align with the real-world experiences the teachers encountered when entering the field (Chelsey & Jordan, 2012). School psychologists serving as consultants bridge the deficit gaps left by teacher training programs and provide essential training to support teachers. When teachers are confident in their work environment, they are more successful and more likely to remain in education (Albright et al., 2017).

Consulting with a school psychologist regarding classroom management issues likely leads to increased teacher confidence as they learn new strategies with the support of a colleague. In a survey of novice teachers, participants identified a desire to increase opportunities to discuss ideas with colleagues and to receive support and guidance without worrying about negative evaluations (Prilleltensky et al., 2016). An effective school consultant builds rapport with teachers, eliminates the fear of retribution, and solidifies the likelihood that teachers will reach out for guidance more often (Kratochwill et al., 2014). Unlike administrators, school psychologists are not in an evaluative role and are better able to build rapport and support teacher training.

What Tools Do Consultants Use?

Consultants use many tools during the consultation process. Any tool used to collect information must be technically sound, providing reliable and valid data (Briesch et al., 2010). Tests, rating scales, functional assessments, and direct observations are commonly used by consultants (Kratochwill et al., 2014). Tests are used to assess student academic or achievement concerns and include curriculum-based assessments (Howell & Hosp, 2014), as well as intelligence and achievement tests (Briesch et al., 2010). Rating scales are often used to assess student behavior or student social-emotional characteristics. Rating scales can also be used to

assess behaviors through performance-based assessments (Steege et al., 2001), for example rating a child's ability to work with others on a scale of 1-5. Functional assessment strategies are used to identify student problem behaviors and determine academic and social competency needs within the ecological context of the classroom (Kratochwill et al., 2014), for example, assessing whether instructions are conveyed explicitly, if goals are appropriate, if the student performed after simple and specific prompts, and whether classroom rules are present and understood (Codding et al., 2014).

Direct Observation. Perhaps the most common tool in consultation and used by school psychologists in general, is direct observation (Hintze, 2005; Steege et al., 2001). Direct observation can be used to assess student behaviors and performance (Briesch et al., 2015), including entry-level skills, as well as teacher practices, such as management strategies (Rosenfield, 2014). When observing a student, a baseline for target behaviors is established and antecedents and consequences can be identified. Further, student skill sets can be observed by comparing direct observation with previous work (Hughes et al., 2014). Observation within the consultative relationship can further allow teachers to voice their concerns and address them jointly with the school psychologist (Rosenfield, 2014). During this process, the school psychologist should encourage the teacher to recognize their own strengths and identify useful supports (Hughes et al., 2014).

Given the frequency at which school psychologists utilize direct observation in consultation and in practice, it is imperative observational tools are psychometrically sound. Measurement, or assigning numbers to behaviors to draw inferences, must be trustworthy. This is especially important when the data is used for important decisions in a child's life, such as placement in special education (Briesch et al., 2010) or assisting a teacher in tailoring classroom

management strategies to positively impact student behavior and academic achievement. The first step in quantifying observable phenomenon is to generate an operational definition for the behavior to be observed (Hintze, 2005). Methods in direct observation should also display adequate levels of interrater, or interobserver, agreement. Systematic direct observation has been shown repeatedly to have strong interrater agreement (Briesch et al., 2010).

Prior to analyzing whether data reflect what they intend to measure, reliability (the consistency between two observers' measurements of the same behavior under the same conditions) and measurement error must be evaluated (Hintze, 2005). Assessing interobserver agreement allows for the quantification of the degree of agreement between multiple coders (observers) making independent ratings about the same individual or group (Hallgren, 2012). Preferably, observers are selected randomly to provide the best estimate of interobserver agreement. Without random selection, however, interobserver agreement indices provide an adequate description of the agreement between the observers actually employed (Hintze, 2005). There are various indices for measuring interobserver agreement. Depending on the intentions of the observation, the analyst can utilize a smaller/larger index, percentage agreement index, occurrence/nonoccurrence agreement indices, the coefficient kappa, and the coefficient phi (Hintze, 2005).

Interobserver agreement analysis is used to determine how much observed score variance is due to true score variance after measurement error has been eliminated. True scores and measurement error cannot be assessed directly, but rather estimated through the covariance among a set of observed scores, allowing for computation of interobserver agreement of a measurement instrument (Hallgren, 2012). Two statistics are commonly used when evaluating interobserver agreement. Cohen's kappa is used for nominal variables and corrects for agreement

that may be expected by chance. Intra-class correlation is applied to ordinal, interval, and ratio variables and incorporated magnitude of disagreements in computing agreement estimates (Hallgren, 2012).

Direct Observation and Existing Studies Examining Psychometrics

Previous studies have analyzed the psychometrics of observational methods and tools commonly used in practice using a variety of methods. Few observational techniques have been psychometrically validated, with researchers utilizing a wide array of methods to measure similar behaviors in the classroom (Briesch et al., 2015). Briesch and colleagues (2015) compared the most frequently used techniques to collect direct observation data in the classroom, such as individual and group. Observations were completed using video footage of classroom instruction. The observation procedures included observing a different student per interval (i.e., observing students one at a time in order as seated), observing groups (i.e., observing one row of students at a time), and observing a few target students individually (i.e., observing only students receiving intervention, as in single-case research). Interobserver agreement was assessed by selecting 33% of the observations at random and having a second observer code. The results of the study showed that observing students in a previously determined order (e.g., beginning with a student in the front and progressing around the desks in order) yielded the most accurate results, however the accuracy across methods of selecting students to observe did not vary significantly (Briesch et al., 2015).

In another study, researchers assessed the reliability of the Behavior-Specific Praise-Observation Tool (BSP-OT) to establish the tool as a reliable observational measure of behavior-specific praise (Markelz et al., 2020). Observations were conducted using 15-minute videos of authentic classroom settings. The researchers used Cohen’s kappa and intra-class correlation to

respectively measure the four raters (experts in behavior-specific praise) to a set of standards and the amount of agreement overall and across categories. Each measure yielded adequate reliability, with high Cohen's kappa scores (at or above 0.8) and a high degree of intra-class correlation (average of 0.78) between raters for all measures (Markelz et al., 2020, p. 8).

The Classroom Measurement Tool (CMOT) and the Five-in-20 Classroom Observation Tool (FCOT; used in the current study) were created recently utilizing the five critical features of classroom management proposed by Simonsen and colleagues (2008). Simonsen and colleagues (2020) created the CMOT and sought to validate the tool in their study. The results of the study showed that the items included on the CMOT were supported by eight experts in the field after review of the tool and had adequate interrater reliability. The CMOT includes ten total items, four of which are likely to happen during instructional activities and six of which may occur less often (Simonsen et al., 2020).

Concurrently, Cardot (2021) created and piloted the FCOT. The FCOT includes each of the five critical features and the 21, evidence-based strategies identified by Simonsen and colleagues (2008). Observers select whether each of the 21 strategies were used by the teacher during a 20-minute observation. If a strategy was observed, the observer rates (1-5) the quality of the strategy (i.e., how closely implementation aligned with the operational definition). The FCOT also includes sections next to each critical feature for additional notes to be taken, as well as an area to note the frequencies of praise and reprimands within the observation.

Anecdotally, the pilot study (Cardot, 2021) found many participants reported the tool to be useful and easy to use, while some believed the 20-minute observation was too short to gather enough information to complete the form. In the pilot study, Cardot (2021) calculated interobserver reliability with eight of the 39 observations (21%). For strategy endorsement,

agreement between raters using Cohen's kappa was 0.58, and for quality ratings agreement between raters was 0.68 (pp. 33-34). In the Cardot (2021) pilot study, interobserver reliability was calculated between two observers among eight different 20-min observations. The present study aims to gain a better estimate of the FCOT's interobserver reliability by asking 40-50 participants to code the same 20-min video and using Cohen's kappa and intra-class correlation to calculate interobserver reliability estimates.

The Current Study

Implementing effective classroom management strategies is crucial to student success and teacher satisfaction and retention. Unfortunately, many teachers leave the field due to issues related to student behavior and classroom management (Freeman et al., 2014), likely stemming from incongruency between teacher training programs and the reality of the classroom (Chelsey & Jordan, 2012). As consultants, school psychologists and others trained to observe and consult with teachers can support and build teachers' skills, to positively impact student behavior and achievement. An observation tool (like the FCOT), which assesses the five critical features of classroom management could help guide school psychologists' consultation; however, it is important to know whether the FCOT is technically sound and feasible to use. Although an initial pilot study (Cardot, 2021) suggested the FCOT may have adequate interobserver reliability, these results were obtained from 8 observations (only 21% of the sample). Therefore, the current study aimed to further assess the reliability of the FCOT by using a single 20-min observation across at least 40 participants. The following research questions guided this study:

1. What is the interrater reliability of the FCOT across observers? Using a single 20-min video, how consistent were observers' ratings with each other?

2. What is the accuracy of the FCOT with a standard coded FCOT (determined by the primary investigator and research mentor)? Using a single 20-min video, how consistent were observers' ratings with a standard coded FCOT?

Method

Participants

Participants for the present study included 44 school professionals who, as a function of their position in a K-12th grade school, conducted teacher consultation and/or student observations. Recruitment efforts took the form of email invitations to EIU alumni, social media postings on school psychologist forums, advertisement on the Illinois School Psychology listserve, and encouragement for participants to recruit other school professionals eligible to participate. These efforts primarily targeted school psychologists, however other school staff expected to provide consultation as a part of their role were also invited to participate. Upon returning completed study materials, each participant was compensated $10 (funding was secured by the PI who was awarded the EIU Graduate Alumni Fund Outstanding Research/Creative Activity Award).

Most participants were White (88%), female (91%), and school psychologists (84%). The remaining non-school psychologist participants (16%) included principals, an instructional coach, a behavior consultant, a special education coordinator, a past school psychologist, and a school psychology graduate student. Most participants worked in Illinois (43%) or Indiana (27%). Two participants each worked in Colorado, Ohio, and Wisconsin, and one participant each worked in Massachusetts, Minnesota, New York, Pennsylvania, Virginia, Washington, and West Virginia.

Work experience (in current role) was evenly distributed between early career (i.e., those with five or fewer years; 45%) and those with six or more years of experience (55%). All but one participant had graduate education with most participants (64%) holding a specialist degree. Most participants had graduate training (i.e., took a graduate course) in consultation (89%), behavior management (89%), and direct observation (80%). Although three participants reported conducting zero observations per month, a little less than half (43%) of participants reported conducting 1-5 observations per month. The remaining participants (50%) conducted 6-20+ observations per month. See Table 1 for additional demographic information.

Table 1

Participant Demographic Characteristics (N = 44)

Characteristics		*N*	%
Gender			
	Female	40	91
	Male	4	9
Race			
	White	39	88
	Black/African American	3	7
	Two or more Races	2	5
Community			
	Rural	21	48
	Suburban	14	32
	Urban	9	20
State			
	Illinois	19	43
	Indiana	12	27
	Other States	13	30
Job Title			
	School Psychologist	37	84
	Principal	2	5
	Instructional Coach	1	2
	Behavior Consultant	1	2
	Other Professional	3	7
Experience			
	<1 year	2	4
	1-5 years	18	41
	6-10 years	10	23

Participant Demographic Characteristics (continued)

	11-20 years	7	16
	20+ years	7	16
Graduate Course			
	Direct Observation		
	Yes	35	80
	No	9	20
	Consultation		
	Yes	39	89
	No	5	11
	Behavior Management		
	Yes	39	89
	No	5	11
Education			
	Bachelors	1	2
	Masters	10	23
	Specialist	28	64
	Doctorate	5	11
Observations per Month			
	0	3	7
	1-5	19	43
	6-10	10	23
	11-15	4	9
	16-20	6	14
	More than 20	2	5

Materials

The materials for the present study included the FCOT, informed consent form, a demographics questionnaire, the link to the observation video, and the standard coded FCOT (see Appendices). All forms excluded identifying information (i.e., names replaced by ID numbers). The demographics form (Appendix A) asked participants to provide their gender, age, race, job title, years of experience, whether they took a behavior management course and consultation course as part of their educational training, and to estimate the number of observations they complete on average per month. The classroom observation video (Appendix

B) is 20 minutes in length and retrieved from YouTube (available for public viewing). The video, titled "2nd Grade Vocabulary Lesson" showed a second-grade teacher conducting full-class vocabulary instruction and an activity. Most of the video entailed large group carpet instruction (55%), with the remainder of the video including large group interactive activity (15%), small group vocab relay race (15%), and transitions to lessons and activities (10%).

The primary investigator and faculty supervisor created a coded copy of the FCOT (Appendix C). This copy was created as an accurately coded FCOT to compare to participants' forms. The primary investigator and faculty supervisor independently watched the video and coded the FCOT, then compared ratings to reach agreement for the combined copy. Disagreements were discussed until complete agreement was reached. Agreement was calculated at $k = .806$, indicating near perfect agreement.

The FCOT (Appendix D) has five critical features (i.e., maximizing structure and predictability, establishing and teaching expectations, engaging students in observable ways, recognizing appropriate behavior using various strategies, and responding to inappropriate behavior using various strategies). Each critical feature has its own associated classroom strategies (range from 2-6). To complete the form, the observer looks for evidence (e.g., teacher demonstrates or physical evidence) of the teacher using any of the strategies during a 20-min observation. If at any time, there is evidence of the strategy, the observer marks "yes," the strategy was observed (i.e., the strategy was endorsed). Next to each strategy there is a quality rating. If the observer indicated "yes," the strategy was observed (i.e., endorsed), they then rate the quality of that strategy (1 = inconsistent with strategy description to 5 = consistent with strategy description). There is also a column where the observer can write notes. Last, throughout the 20-min observation, the observer records the frequency of the teachers' use of

praise (behavior-specific and general) and reprimand using operational definitions provided. The total strategy score is obtained by summing the number of "yes" endorsements (total possible strategy score range is 0 - 21). The total quality score is obtained by summing the 0-5 strategy ratings (total possible quality score range is 0 – 105). A total quality score of "0" indicates the observer did not observe any of the 21 strategies and the Total strategy score is also "0". Praise and reprimand tallies are totaled to obtain a total behavior-specific praise, total general praise, and total reprimand. However, for this study praise and reprimand tallies were not analyzed.

Procedures

IRB approval was obtained (IRB #21-134). Participants were recruited through email invitations to EIU alumni, social media postings on school psychologist forums, advertisement on the Illinois School Psychology listserve, and encouragement for participants to recruit other school professionals eligible to participate. When a participant indicated they were interested in participating, the investigator emailed them the following: (a) Informed Consent, (b) Demographics Questionnaire, (c) Five in 20 Observation Form (FCOT), and (d) link to the observation Video. The investigator followed up with participants via email or phone call to review the FCOT and answer any questions. Step-by-Step Directions for how to complete the FCOT were located at the top of the observation form. Participants were directed to view the video and conduct the "observation" by completing the FCOT while watching the video. Participant FCOT data was compared to the standard coded FCOT completed by the investigator and faculty mentor. Reliability between participants and reliability between participants' FCOT and the standard coded FCOT were analyzed.

Analytic Plan

The first research question: What is the interrater reliability of the FCOT across observers? Using a single 20-min video, how consistent were observers' ratings with each other? was answered by examining intraclass correlation (ICC) and the alpha coefficient. These statistical methods were used to evaluate the degree of reliability between the observers and the extent to which the items were consistently scored. Intraclass correlation provides a magnitude of agreement between the participant's ratings (Hallgren, 2012). Intraclass correlation may be conducted using absolute agreement or consistency, as well as single measures or average measures. Absolute agreement examines raters that assign the same score for the same item, whereas consistency compares ratings in an additive manner (Koo & Li, 2016). Intraclass correlation of single measures examines the reliability of ratings for a single, typical rater. Single measures, representing a more rigorous assessment of reliability, tends to yield lower values than average measures. Intraclass correlation of average measures measures the reliability of several raters averaged together (Schoonjans, 2021).

Intraclass correlation was calculated for both strategy endorsement and quality rating. Participants may have rated the quality of a strategy from 0-5, with scores of zero representing items marked as not observed. The alpha coefficient provided an overall measure of how consistently ratings were assigned to each item on the FCOT across participants (Feldt et al., 1987). The intraclass correlation coefficient and alpha coefficient were calculated using IBM SPSS Version 29. on a computer utilizing Windows 10.

The second research question: What is the reliability of the FCOT with a standard coded FCOT (determined by the primary investigator and research mentor)? Using a single 20-min video, how accurate were observers' ratings with a standard coded FCOT? was answered using

Cohen's Kappa (Sim & Wright, 2005) and Percent Agreement (Watkins & Pacheco, 2000). These statistical measures were used to assess the accuracy of the participant codes with the standard coded FCOT created by the primary investigator and the research mentor. Cohen's kappa provided an estimate of accuracy corrected for accuracy that would happen by chance. An average was then calculated by adding each kappa value and dividing by the total number of kappa values to represent an overall index of agreement (Hallgren, 2012). Weighted kappa was used as data were ordinal, where a participant may rate each item on a 0-5 scale.

Percent Agreement provided an estimate of agreement between participant forms and the standard coded form. A percentage was calculated for each participant for whether they endorsed each strategy (i.e., Yes or No rating) and their quality strategy rating. Percent agreement was first calculated for each participant, individually, by dividing the number of matching endorsements by the total number of items (21) and multiplying the value by 100 (Watkins & Pacheco, 2000). This yielded the percent agreement of each participant to the standard coded FCOT in terms of strategy endorsement. Values 75% and above are considered acceptable, while values closer to 90% are preferred (Graham et al., 2012). Percent agreement was similarly calculated for each participant in terms of quality ratings.

Results

Reliability Across Observers

Total Strategy Score

The total strategy score represents the number of "yes" (i.e., the strategy was observed) endorsements and ranges from 0-21. The intraclass correlation coefficient was used to analyze the consistency of observers' "yes" versus "no" endorsements. In computing intraclass correlation for the dataset, one item (Planned Ignoring) was excluded due to a missing datum. Absolute agreement was used to analyze consistency of strategy, as there were only two potential ratings (yes = 1, no = 0). Values for absolute single measures and absolute average measures are reported.

Intraclass correlation coefficients less than 0.5 indicate poor reliability, 0.5-0.75 indicate moderate reliability, 0.75-0.9 indicate good reliability, and greater than 0.9 indicate excellent reliability (Koo & Li, 2016). The resulting intraclass correlation coefficient of average measures was *ICC* = .99, whereas the ICC of single measures was *ICC* = .67. The average measures ICC indicated excellent reliability, whereas the single measures ICC indicated moderate reliability. Cronbach's alpha ($\alpha = .99$), indicated strong reliability (Glen, 2023).

Total Quality Score

The total quality score represents the actual ratings given by participants (i.e., 1 = inconsistent with strategy description to 5 = consistent with strategy description). Ratings of 0 were included for participants who indicated "no," the strategy was not observed. The intraclass correlation coefficient was used to analyze the consistency of observers' ratings. In computing intraclass correlation for the dataset, one item (Planned Ignoring) was excluded due to a missing datum. Absolute agreement and consistency were both conducted, as quality ratings could range

from 0-5, allowing for differing values in absolute agreement and consistency. The resulting ICC of single measures using a consistency model indicated moderate reliability (*ICC* = .67). The ICC of single measures using an absolute agreement model also indicated moderate agreement (*ICC* = .66). The ICC of average measures was equal in the consistency and absolute agreement models and indicated excellent reliability (*ICC* = .99). Cronbach's alpha (α = .99) also indicated strong reliability.

Accuracy with Standard Coded FCOT

Cohen's Kappa

Kappa values of less than zero indicates poor agreement, values between .01-.20 indicate slight agreement, values between .21-.40 indicate fair agreement, values between .41-.60 indicate moderate agreement, values between .61-.80 indicate substantial agreement, and values between .81-1 indicate almost perfect agreement (Sim & Wright, 2005). The average weighted kappa for strategy endorsement (M_k = .71; SD_k = .14) indicated substantial agreement, while the average weighted kappa for quality rating (M_k = .56; SD_k = .11) indicated moderate agreement overall. Sixteen participants (36%) had substantial agreement with the standard coded FCOT; 26 participants (59%) had moderate agreement; one participant (2%) had fair agreement; and one participant (2%) had slight agreement.

Percent Agreement

Overall Endorsement and Quality Rating. Percent agreement was also used to compare each participant's ratings with the standard coded FCOT. The averages of the yielded percentages were calculated to yield values representative of the sample. The average percent agreement for endorsed strategies (i.e., whether a strategy was observed in the video) was 85.37% (*SD* = 6.92%; range 76.19%-100%). The average percent agreement for participant

quality ratings was 60.24% (*SD* = 11.02%; range 33.33%-76.19%). The percent agreement for strategy endorsement was considered acceptable, while the percent agreement for quality rating was low.

Critical Feature Endorsement and Quality Rating. Percent agreement was also calculated by strategy to investigate which items were rated more consistently by the observers. This was done by summing the number of ratings within a strategy that match the standard coded FCOT and dividing by the total number of ratings. The average percent agreement for strategy endorsement was 85.35% (*SD* = 18.57%) while the average percent agreement for quality rating across items was 60.35% (SD = 29.10%). The average percent agreements for items were used to investigate percent agreement within each Critical Feature. This was done by summing the percent agreements for the strategies within a Critical Feature and dividing by the total number of strategies within that feature. In looking at each critical feature, Critical Feature 4 had the highest average percent agreement across items, whereas Critical Feature 2 had the lowest average percent agreement across items. These data are presented in Table 2.

Critical Feature 1: Maximizing Structure & Predictability. Critical Feature 1 includes four classroom management strategies. For strategy endorsement, Easy Traffic Flow, Structured, and Rules posted had percent agreement values at or above 90%. Schedule posted had the lowest percent agreement value (84.09%). For quality rating, Structured and Schedule Posted had percent agreement values at 81.82% and 84.09%, respectively. Easy Traffic Flow (38.64%) and Rules Posted (38.64%) were much lower. The average percent agreement of strategy endorsement was 92.62% (*SD* = 5.65%), while the average for quality rating was 60.80% (*SD* = 22.17%).

Critical Feature 2: Establishing and Teaching Expectations. Critical Feature 2 includes two classroom management strategies (Rules: Taught and Reviewed and Active Supervision). Each had high percent agreement values (100% and 95.45%, respectively). For strategy endorsement, the percent agreement values were as follows: Rules: Taught and Reviewed (100%) and Active Supervision (95.45%). Percent agreement for quality rating was much lower (45.45% and 31.82%, respectively). The average percent agreement of strategy endorsement was 97.73% (*SD* = 2.27%), while the average for quality rating was 77.27% (*SD* = 6.82%).

Critical Feature 3: Engaging Students in Observable Ways. Critical Feature 3 includes five classroom management strategies, four (Opportunities to respond, Direct Instruction, Computer Assisted Instruction, Guided Notes) had percent agreement values above 88%. Percent agreement for Class-Wide Tutoring was 68.18%. For quality rating, two strategies had percent agreement values above 88% (i.e., Computer Assisted Instruction and Guided Notes). Opportunities to Respond and Direct Instruction had lower percent agreement values for quality ratings (63.64% and 65.91%, respectively). Class-Wide Tutoring (29.55%) had the lowest percent agreement values for quality ratings. The average percent agreement of strategy endorsement was 91.36% (*SD* = 12.40%), while the average for quality rating was 69.55% (*SD* = 24.25%).

Critical Feature 4: Recognizing Appropriate Behavior using Various Strategies. Critical Feature 4 includes four strategies, which all had percent agreement values at or above 90%. For-quality ratings, Token Economies, Class-Wide Group Contingencies, and Behavior Contracts had percent agreement values at or above 90%. The percent agreement for the quality rating for Behavior-Specific Praise was much lower (22.73%). The average percent agreement

for strategy endorsement was 95.46% (*SD* = 3.59%), while the average for quality rating was 77.84% (*SD* = 32.00%).

Critical Feature 5: Responding to Inappropriate Behavior using Various Strategies. Critical Feature 5 includes six classroom management strategies. Response Cost had the highest percent agreement value for quality rating (90.91%), followed by Brief Instructional Corrections for Inappropriate Behavior (79.55%), Performance Feedback (72.73%), Planned Ignoring (60.47%), Differential Reinforcement (59.09%), and Time Out for Reinforcement (25.00%). For quality rating, the percent agreement values were highest for Response Cost (90.91%), Performance Feedback (72.73%), and Planned Ignoring (60.47%). Corrections for Inappropriate Behavior (29.55%) and Time Out from Reinforcement (4.55%) were much lower. The average percent agreement for strategy endorsement was 64.63% (*SD* = 20.81%), while the average for quality rating was 47.96% (SD = 29.38%).

Table 2

Percent Agreement by Item

	Yes/No Agreement	Rating Agreement
Critical Feature 1: Maximizing Structure & Predictability		
Easy Traffic Flow	97.73%	38.64%
Structured	97.73%	81.82%
Rules Posted	90.91%	38.64%
Schedule Posted	84.09%	84.09%
M	**92.62%**	**60.80%**
SD	**5.65%**	**22.17%**
Critical Feature 2: Establishing and Teaching Expectations		
Rules: Taught & Reviewed	100.00%	45.45%
Active Supervision	95.45%	31.82%
M	**97.73%**	**38.64%**
SD	**2.27%**	**6.82%**
Critical Feature 3: Engaging Students in Observable Ways		
Opportunities to Respond (OTR)	100.00%	63.64%

Percent Agreement by Item (continued)

Direct Instruction	100.00%	65.91%
Class-Wide Tutoring	68.18%	29.55%
Computer Assisted Instruction	100.00%	100.00%
Guided Notes	88.64%	88.64%
M	**91.36%**	**69.55%**
SD	**12.40%**	**24.25%**
Critical Feature 4: Recognizing Appropriate Behavior using Various Strategies		
Using Behavior-Specific Praise	93.18%	22.73%
Token Economies	97.73%	97.73%
Class-Wide Group Contingencies	90.91%	90.91%
Behavior Contracts	100.00%	100.00%
M	**95.46%**	**77.84%**
SD	**3.59%**	**32.00%**
Critical Feature 5: Responding to Inappropriate Behavior using Various Strategies		
Brief Instructional Corrections for Inappropriate Behavior	79.55%	29.55%
Performance Feedback	72.73%	72.73%
Planned Ignoring	60.47%	60.47%
Differential Reinforcement	59.09%	29.55%
Response Cost	90.91%	90.91%
Time Out from Reinforcement	25.00%	4.55%
M	**64.62%**	**47.96%**
SD	**20.81%**	**29.38%**
Overall *M*	85.35%	60.35%
Overall *SD*	18.57%	29.10%

Additional Analyses: Participant Demographics

The data were further analyzed based on job title, years of experience, degree type, number of observations conducted in one month, and whether the participant received instruction in behavior management, consultation, and/or direct observation. These analyses were conducted to examine potential future questions related to the tool's use with different professionals.

Job title

Across school psychologists (n = 37) average kappa indicated substantial agreement for strategy endorsement (M_k = .70; SD = .14) and moderate agreement for quality rating (M_k =.55; .11). Percent agreement for strategy endorsement and quality ratings were 85.05% and 59.28%, respectively. The reliability estimates based on job title are presented in Table 3.

Table 3

Reliability Based on Job Title

		Mean Kappa (Standard Deviation)		Mean Percent Agreement (Standard Deviation)	
	N	Strategy	Quality	Strategy	Quality
School Psychologist	37	.70 (.14)	.55 (.11)	85.05 (6.93)	59.28 (11.21)
Principal	2	.91 (.09)	.70 (.05)	95.24 (4.76)	64.29 (2.38)
Instructional Coach	1	.81	.65	90.48	66.67
Behavior Consultant	1	.61	.66	80.95	71.43
Other	3	.65 (.05)	.55 (.10)	82.54 (2.24)	63.49 (11.88)

Experience

The greatest agreement based on years of experience was observed for participants reporting 11 to 15 years of experience (N = 2), while the lowest agreement was observed for participants reporting less than 5 years of experience (N = 15). These data are reported in Table 4. There was little to no correlation between years of experience and agreement in strategy endorsement (r = -.09) and quality rating (r = .19).

Table 4

Reliability Based on Years of Experience

		Mean Kappa (Standard Deviation)		Mean Percent Agreement (Standard Deviation)	
	N	Strategy	Quality	Strategy	Quality
Less than 5	15	.69 (.14)	.53 (.12)	84.76 (7.62)	58.73 (12.13)
5 to 10	15	.71 (.13)	.57 (.09)	85.71 (6.27)	57.46 (9.91)
11 to15	2	.90 (0.0)	.68 (.03)	95.24 (0.0)	66.67 (0.0)
16 to 20	5	.67 (.11)	.54 (.10)	83.81 (4.86)	62.86 (12.92)
More than 20	7	.68 (.14)	.60 (.09)	84.25 (6.58)	65.71 (7.13)

Degree Type

The data were also sorted and analyzed based on the participants' highest obtained degree. The greatest agreement was observed in participants with a doctorate (n = 5), while the lowest agreement was observed in participants with a masters (n = 10). These data are presented in Table 5.

Table 5

Reliability Based on Highest Obtained Degree

		Mean Kappa (Standard Deviation)		Mean Percent Agreement (Standard Deviation)	
	N	Strategy	Quality	Strategy	Quality
Bachelors	1	.61	.57	80.95	66.67
Masters	9	.70 (.17)	.54 (.16)	84.76 (8.73)	57.14 (9.99)
Specialist	28	.73 (.13)	.56 (.10)	86.54 (6.12)	59.97 (11.68)
Doctorate	5	.61 (.11)	.58 (.07)	80.95 (5.22)	66.67 (6.02)

Observations per Month

For reported number of monthly observations, participants reporting conducting 16-20 observations per month (N = 5) had the greatest reliability by quality rating, while those conducting 5 to 10 observations per month (N = 12) had the greatest reliability by strategy endorsement. Those reporting conducting 5 to 10 observations per month (N = 12) had the lowest reliability for quality rating, while those reporting conducting less than five observations per month (N = 20) had the lowest reliability for strategy endorsement. These data are reported in Table 6. There was little to no correlation between number of observations per month and agreement for strategy endorsement (r = .05) nor quality rating (r = .05).

Table 6

Reliability Based on Number of Observations Conducted in Practice

		Mean Kappa (Standard Deviation)		Mean Percent Agreement (Standard Deviation)	
	N	Strategy	Quality	Strategy	Quality
Less than 5	20	.69 (.14)	.56 (.10)	84.76 (6.67)	61.19 (10.46)
5 to 10	12	.73 (.16)	.54 (.14)	86.45 (8.44)	59.76 (10.71)
11 to 15	4	.71 (.14)	.55 (.03)	85.71 (6.73)	52.38 (12.14)
16 to 20	5	.69 (.10)	.60 (.10)	84.76 (4.67)	60.95 (11.02)
More than 20	3	.71 (.08)	.59 (.06)	85.71 (3.89)	65.08 (8.98)

Graduate Coursework

The data were also analyzed based on enrollment in classes involving instruction in behavior management, consultation, and direct observation. The reliability estimates across topics and between participants indicating yes versus no were roughly equivalent. The data are presented in Table 7.

Table 7

Reliability Based on Class Enrollment

		Mean Kappa (Standard Deviation)		Mean Percent Agreement (Standard Deviation	
	N	Strategy	Quality	Strategy	Quality
Behavior Management					
Yes	39	.70 (.13)	.55 (.11)	84.96 (6.70)	60.02 (11.34)
No	5	.77 (.16)	.62 (.12)	88.57 (7.74)	61.90 (7.97)
Consultation Course					
Yes	39	.70 (.14)	.55 (.11)	85.09 (6.83)	59.90 (11.29)
No	5	.75 (.14)	.62 (.11)	87.62 (7.13)	62.86 (8.20)
Direct Observation					
Yes	35	.72 (.14)	.56 (.10)	86.10 (6.68)	60.08 (10.18)
No	9	.65 (.13)	.58 (.15)	82.54 (7.10)	60.85 (13.80)

Discussion

There is a need for a psychometrically sound observation tool to support teachers' acquirement and use of empirically supported classroom management practices. Effective classroom management practices are crucial to fostering a positive learning environment, student achievement, and student social-emotional development, as well as teacher job satisfaction. Simonsen and colleagues (2008) conducted a review of classroom management strategies and found 21 strategies with empirical support for their effectiveness, grouped into five "critical features." These critical features and strategies, paired with the idea that direct observation and consultation are crucial in supporting teachers, serve as the basis of the Five-in-20 Classroom Observation Tool (FCOT).

The present study sought to assess the reliability of the Five-in-20 Classroom Observation Tool (FCOT). This study builds on the pilot study, where interobserver reliability of strategy endorsement (k = .676; substantial) and of quality rating (k = .580; moderate) was measured in eight pairs of raters from a larger sample (n = 39). The current study included 44 school professionals who, as a function of their position in a K-12th grade school, conduct teacher consultation and/or student observations. Most participants were White, female, school psychologists aged 30-39. The sample included participants who worked in 12 states from all four geographic regions (West, Midwest, South, Northeast) of the United States. Contrary to the pilot study, where agreement was greater for quality rating than for strategy endorsement, the present study yielded greater agreement for strategy endorsement (k = .705; substantial) than quality rating (k = .561).

First Research Question

The first research question (What is the reliability of the FCOT across observers? Using a single 20-min video, how consistent are observers' ratings with each other?) was investigated using intraclass correlation and the alpha coefficient. Intraclass correlation using absolute agreement was used to analyze agreement of participants' yes/no ratings (0-1). This yielded excellent reliability of average measures and moderate reliability of single measures. The yielded Cronbach's alpha for yes/no ratings suggested strong reliability. These results indicate that participants mostly agreed on whether a given strategy was used by the teacher in the video.

Intraclass correlation using consistency and absolute agreement measures was used to analyze agreement of participants' quality ratings of the strategies (0-5). Both absolute agreement and consistency yielded moderate reliability of single measures and excellent reliability of average measures. The yielded Cronbach's alpha for quality ratings suggested

strong reliability. These results indicate that participants displayed moderate agreement on actual quality ratings of each FCOT strategy. Absolute agreement was marginally higher for yes/no agreement (ICC = .669) than for quality rating agreement (ICC = .664).

The literature does not specify criteria for an acceptable level of intraclass correlation (Koo & Li, 2016; Volpe et al., 2005). One study examining the reliability of an observation tool indicated intraclass correlation of single measures to be acceptable at ICC = .58 and intraclass correlation of average measures to be acceptable at ICC = .73 (Reed & Edelbrock, 1983). A more recent study examining the reliability of an observation tool used an ICC cutoff of ICC = .75 (Markelz et al., 2020). Based on this assumption, the results derived from the intraclass correlation coefficients yielded from the present study may indicate acceptable consistency across raters.

Second Research Question

The second research question (What is the accuracy of the FCOT with a standard coded FCOT (determined by the primary investigator and research mentor)? Using a single 20-min video, how consistent are observers' ratings with a standard coded FCOT?) was investigated using Cohen's kappa and percent agreement. Weighted kappa was calculated comparing each individual participant's responses to the standard coded FCOT. The average of the kappa values was then calculated to yield an overall reliability estimate, which was found to be moderate. Most participants (59%, n = 26) had moderate agreement with the standard coded copy, while 36% (n = 16) had substantial agreement. Cohen's kappa suggested participants' interobserver reliability with the standard coded FCOT was substantial for strategy endorsement and moderate for quality rating.

Percent agreement was also used to compare each participant's ratings with the standard coded FCOT. An average percent agreement was obtained to identify a value representative of the participant sample. The average percent agreement for strategy endorsement was 85.37%, while the average percent agreement for quality ratings was 60.24%. These results suggest that participants were more reliable in indicating whether a strategy was used in the video than rating the quality of the strategy on a scale of 0-5.

Percent agreement was also used to see how reliable individual items were on the FCOT. Critical Feature 4: Recognizing Appropriate Behavior using Various Strategies yielded the greatest average percent agreement across its strategies. Three of the four strategies included in Critical Feature 4 were not displayed during the video, which likely increased agreement due to the elimination of needing to rate those items in participants who accurately indicated "no." Critical Feature 2: Establishing and Teaching Expectations yielded the lowest average percent agreement across its strategies. This critical feature included only two strategies, both of which were displayed during the video, likely impacting the percent agreement magnitude when compared to a critical feature with strategies not displayed.

Additional Analyses

Additional analyses were conducted examining job title, experience, degree type, observations per month, and graduate coursework. However, caution should be used when comparing these reliability estimates. For instance, the magnitude of differences may be exaggerated due to differences in sample size within each analysis. School psychologists, the primary target of this investigation, had moderate agreement with the standard coded FCOT. Principals, instructional coaches, and behavior consultants displayed substantial agreement with

the standard coded FCOT. With more participants identifying as school psychologists, there is more room for variability and outliers in this subsample (n = 37).

Participants with more experience (i.e., 11 to 15 years of experience) yielded greater agreement than those with 10 or less years of experience, while early career professionals (less than 5 years of experience) yielded the lowest agreement. In addition to those with 11-15 years of experience, participants with more than 20 years of experience also displayed substantial agreement. There was little to no correlation between years of experience and level of agreement based on kappa. Similarly, participants with a specialist degree had the greatest agreement for strategy endorsement, and those with a doctorate had the greatest agreement for quality rating. There may be a positive relationship between more education and stronger reliability.

There were three areas of graduate coursework examined in this study: behavior management, consultation, and direct observation. Participants indicated on the demographics questionnaire whether they had received instruction in these areas. The mean kappa values for those who did receive instruction were close in value to those who did not receive instruction across course topics.

Limitations

The present study is the first to examine the psychometric properties of the FCOT. However, there are limitations to note. First, this study was limited in part using a video. For example, although participants were instructed to watch the video once, some may have taken advantage of the opportunity to rewind and rewatch. Rewinding and rewatching has the potential to improve participant's reliability. Watching a video is also different from the "real world setting" where rewinding and rewatching is not possible. It is impossible to know whether participants only watched the video once. On the other hand, coding a video is advantageous for

this type of study as the content within the video (i.e., strategies used) remain the same. In other words, each participant observed the same teacher give the same lesson, using the same strategies. It would not be practical to have 40+ observers within the same classroom at the same time. Future studies might examine the FCOT live by utilizing live-stream or video call technology. Also, some participants reported technical difficulties that could have created issues with accurate rating, such as video skipping and wi-fi connectivity. It is possible that this impacted the accuracy of reliability estimates. To overcome this limitation, future studies might show the video in a controlled setting, where the primary investigator monitors/resolves technical issues and ensures participants only view the video once without rewinding.

Another potential limitation was the lack of training for participants in using the FCOT. Participants were asked to review the FCOT (i.e., read strategy descriptions, instructions on how to use the tool, etc.) prior to coding the observation video. However, whether participants reviewed the FCOT prior to coding could not be ensured. In the future, investigators may opt to review the FCOT with participants when disseminating the materials, such as via phone call or video call. A manual may also be developed. Having a manual may increase the reliability of observers as it would describe the tool and its features in greater detail than that included in the FCOT.

Implications and Future Directions

Preliminary evidence suggests that the FCOT may be a reliable tool to use in direct observation of teachers' use of evidence-based classroom management strategies. To improve upon these reliability estimates, alterations may be needed with the tool or its methods. For example, training may be necessary to increase interobserver agreement. The training could involve an in-depth description of each strategy and an explanation of the 1-5 quality ratings

with examples of what might constitute differences between ratings (e.g., 2 rating vs. 4 rating). A future study may investigate the difference in interobserver agreement between those receiving brief training in the FCOT and those receiving none. If such a study yields no difference between groups, alterations to the tool itself may be explored to increase rater accuracy and agreement. For instance, certain critical features or strategies with lower reliability may be removed. Some strategies on the tool may be difficult to observe or may not occur regularly in the classroom setting. A future study may investigate the reliability of the tool after removing such items. Future studies may also investigate whether requiring participants to reach a training criterion (e.g., 90% agreement on each critical feature) prior to coding the video.

Conclusion

There remains a need in the field for a reliable classroom observation tool to assist in recommending areas of improvement for teachers' classroom management practices. Few direct observation tools examining student behavior have been psychometrically examined and even fewer direct observation tools examining teacher behavior. The Five-in-20 Classroom Observation Tool shows promise for reliably assessing teachers' use of evidence-based classroom management strategies. Therefore, this tool may be helpful in consultation related to identifying and increasing teachers' use of these strategies. Furthermore, there seems to be a need for a tool like this as participants in this study reached out to the PI after using the FCOT with positive feedback about the tool and questions about further use and distribution.

References

Albright, J. L., Safer, L. A., Sims, P. A., Tagaris, A., Glasgow, D., Sekulich, K. M., & Zaharis, M. C. (2017). What factors impact why novice middle school teachers in a large midwestern urban school district leave after their initial year of teaching. *NCPEA International Journal of Educational Leadership Preparation, 12*(1).

Aloe, A. M., Amo, L. C., & Shanahan, M. E. (2014). Classroom management self-efficacy and burnout: A multivariate meta-analysis. *Educational Psychology Review, 26*, 101-126. https://doi.org/10.1007/s10648-013-9244-0.

Bahr, M. W., Leduc, J. D., Hild, M. A., Davis, S. E., Summers, J. K., & McNeal, B. (2017). Evidence for the expanding role of consultation in the practice of school psychologists. *Psychology in the Schools, 54*(6), 581-595. https://doi.org/10.1002/pits.22020.

Briesch, A. M., Chafouleas, S. M., & RIley-Tillman, T. C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systematic direct observation and direct behavior rating. *School Psychology Review, 39*(3), 408-421. https://doi.org/10.1080/02796015.2010.12087761.

Briesch, A. M., Hemphill, E. M., Volpe, R. J., & Daniels, B. (2015). An evaluation of observational methods for measuring response to classwide intervention. *School Psychology Quarterly, 30*(1), 37-49. https://doi.org/10.1037/spq0000065.

Cardot, A. “Five in 20: An exploratory study to develop and pilot an observation tool used to assess the five features critical to effective classroom management” (2021). Masters Theses. 4875. https://thekeep.eiu.edu/theses/4875.

Chelsey, G. M. & Jordan, J. (2012). What’s missing from teacher prep? *Educational Leadership, 69*(8).

Chuang, C., Reinke, W. M., & Herman, K. C. (2020). Effects of a universal classroom management teacher training program on elementary children with aggressive behaviors. *School Psychology Quarterly, 35*(2), 128-136. https://doi.org/10.1037/spq0000351.

Codding, R. S., Sanetti, L. M. H., & Reed, F. D. D. (2014). Best practices in facilitating consultation and collaboration with teachers and administrators. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 525-540). National Association of School Psychologists.

Domitrovich, C. E., Bradshaw, C. P., Berg, J. K., Pas, E. T., Becker, K. D., Musci, R., Embry, D. D., & Ialongo, N. (2016). How do school-based prevention programs impact teachers? Findings from a randomized trial of an integrated classroom management and social -emotional program. *Prevention Science, 17*, 325-337. https://doi.org/10.1007/s11121-015-0618-z.

Erchul, W. P. & Martens, B. K. (2010). *School consultation: Conceptual and Empirical Bases of Practice*. Springer.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*(1), 93-103. https://doi.org/10.1177/014662168701100107.

Freeman, J., Simonsen, B., Briere, D. E., & MacSuga-Gage, A. S. (2014). Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs. *Teacher Education and Special Education, 37*(2), 106-120. https://doi.org/10.1177/0888406413507002.

Fruth, J. D. (2014). Impact of a universal prevention strategy on reading and behavioral outcomes. *Reading Improvement, 51*(3), 281-290.

Garwood, J. D., Vernon-Feagans, L., & the Family Life Project Key Investigators (2017). Classroom management affects literacy development of students with emotional and behavioral disorders. *Exceptional Children, 83*(2), 123-142. https://doi.org/10.1177/0014402916651846.

Glen, S. (2023, March 3). *Cronbach's alpha: Definition, interpretation, SPSS*. Statistics How To. Retrieved April 9, 2023, from https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/cronbachs-alpha-spss/

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative for Psychology, 8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*(4), 507-519. https://doi.org/10.1080/02796015.2005.12088012.

Hintze, J. M. & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*(2), 258-270. https://doi.org/10.1080/02796015.2004.12086247.

Howell, K. W. & Hosp, J. L. (2014). Best practices in curriculum-based assessment. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 159-170). National Association of School Psychologists.

Hughes, T. L., Kolbert, J. B., & Crothers, L. M. (2014). Best practices in behavioral/ecological consultation. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology:*

Data-based and collaborative decision making (pp. 483-492). National Association of School Psychologists.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38*(3), 499-534. https://doi.org/10.3102/00028312038003499.

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kratochwill, T. R., Altshaefl, M. R., & Bice-Urbach, B. (2014). Best practices in school-based problem-solving consultation: Applications in prevention and intervention systems. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 461-482). National Association of School Psychologists.

Lekwa, A. J., Reddy, L. A., & Shernoff, E. S. (2018). Measuring teacher practices and student academic engagement: A convergent validity study. *School Psychology Quarterly, 34*(1), 109-118. https://doi.org/10.1037/spq0000268.

MacSuga, A. S. & Simonsen, B. (2011). Increasing teachers' use of evidence-based classroom management strategies through consultation: Overview and case studies. *Beyond Behavior, 20*(2), 4-12.

Markelz, A. M., Riden, B. S., Zoder-Martell, K. A., Miller, J. E., & Bolinger, S. J. (2020). Reliability assessment of an observation tool to measure praise characteristics. *Journal of*

Positive Behavior Interventions, 23(1), 17-29. https://doi.org/10.1177/1098300720907988.

Oliver, R. M. & Reschly, D. J. (2010). Special education teacher preparation in classroom management: Implications for students with emotional and behavioral disorders. *Behavioral Disorders, 35*(3), 188-199. https://doi.org/10.1177/019874291003500301.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom Assessment Scoring System™: Manual K-3. Paul H Brookes Publishing.

Prilleltensky, I., Neff, M., & Bessell, A. (2016). Teacher stress: What it is, why it's important, how it can be alleviated. *Theory Into Practice, 55*, 104-111. https://doi.org/10.1080/00405841.2016.1148986.

Reed, M. L. & Edelbrock, C. (1983). Reliability and validity of the direct observation form of the child behavior checklist. *Journal of Abnormal Child Psychology, 11*(4), 521-530.

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., & Goel, N. (2011). Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers. *School Psychology Quarterly, 26*(1), 1-13. https://doi.org/10.1037/a0022714.

Reinke, W. M., Stormont, M., Herman, K. C., & Dong, N. (2021). The incredible years teacher classroom management program: Effects for students receiving special education services. *Remedial and Special Education, 42*(1), 7-17. https://doi.org/10.1177/0741932520937442.

Rosenfield, S. (2014). Best practice in instructional consultation and instructional consultation teams. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 509-524). National Association of School Psychologists.

Schoonjans, F. (2021, August 20). *Intraclass correlation coefficient.* MedCalc. Retrieved April 20, 2023, from https://www.medcalc.org/manual/intraclass-correlation-coefficient.php#:~:text=Single%20measures%3A%20this%20ICC%20is%20an%20index%20for,is%20always%20h.

Sim, J. & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268. https://doi.org/10.1093/ptj/85.3.257.

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children, 31*(3). https://doi.org/10.1353/etc.0.0007.

Simonsen, B., Freeman, J., Kooken, J., Dooley, K., Gambino, A. J., Wilkinson, S., VanLone, J., Walters, S., Byun, S. G., Xu, X., Lupo, K., & Kern, L. (2020). Initial validation of the classroom management observation tool (CMOT). *School Psychology Quarterly, 35*(3), 179-192. https://doi.org/10.1037/spq0000357.

Steege, M. W., Davin, T., & Hathaway, M. (2001). Reliability and accuracy of a performance-based behavioral recording procedure. *School Psychology Review, 30*(2), 252-261. https://doi.org/10.1080/02796015.2001.12086113.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339-355. https://doi.org/10.1177/0022487111404241.

Sugai, G. & Horner, R. H. (2002). Introduction to the special series on positive behavior support in schools. *Journal of Emotional and Behavioral Disorders, 10*(3), 130-135. https://doi.org/10.1177/10634266020100030101.

Volpe, R. J., Diperna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review, 34*(4), 454-474.

Wilson, M. S. & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25*(1), 9-23. https://doi.org/10.1080/02796015.1996.12085799.

Appendix A: Demographics Questionnaire

1. Please indicate your gender identity (circle all that apply):

   Male

   Female

   Non-binary/non-conforming

   Transgender

   Prefer not to answer

   _____ (fill-in answer)

2. Please indicate your race/ethnicity

   American Indian or Alaska Native

   Asian

   Black or African American

   Native Hawaiian or Other Pacific Islander

   Hispanic/Latinx

   White

   Two or more races (please specify) _____

   I prefer not to answer

3. Please indicate your age. _____

4. Please list your job title. _____

5. In what state do you work? _____

6. How would you describe the community in which you work? (circle):

   Rural Urban Suburban

7. How many years of experience do you have _____ years.

8. What is your highest obtained degree? (circle):

   Bachelors

   Masters

   Specialist

   Doctorate

   Other (please specify): _____

9. Have you taken an undergraduate or graduate course that focuses on managing student behavior?

   Yes (provide the name of the course, if possible) _____

   No

10. Have you taken a graduate Consultation course?

11. Have you taken a graduate course where you were trained in direct observation?

12. How many direct observations do you do in a month?

Appendix B: Observation Video Link

https://youtu.be/8R1hy3uHds0

Appendix C: Standard Coded FCOT

Observer Code: Master Copy Job Title: ________ Reason for the Observation: ________ (e.g., re-eval, consultation).

Reliability Observer Code: ________ Job Title: ________ Date: ________ School Code: ________ Observation Length: 20 min

Teacher Code: ________ Class Description 2nd grade (e.g., 5th grade general ed). How many students in the classroom ________ ?

Directions 1) **STRATEGY:** Indicate if the classroom strategy was observed during the 20-min observation.

2) **QUALITY:** If a strategy was observed, indicate the quality of the observed strategy (1 = poor, 5 = great).

3) **FREQUENCY:** Using the operational definitions, tally the frequency of Praise and Reprimands observed during the 20-min observation.

	Critical Feature	Classroom Strategy & Description	Observed Yes or No	Quality (Inconsistent with Description – Consistent with Description)	Comments/Notes
1	Maximizing Structure & Predictability	• Easy Traffic Flow (physical arrangement of the room (e.g., adult can easily move in the isle & btw furniture, can access all of room, no roadblocks, no visual obstructions).	☑ Yes ☐ No	1 2 3 4 (5)	
		• Structured (teacher-imposed organization, lesson plan is clear, students are aware of current classroom objectives; visual aids are informative and not distracting)	☑ Yes ☐ No	1 2 3 4 (5)	
		• Rules Posted (classroom rules are visible, positively stated, large enough to read, succinct, linked to SW expectations).	☑ Yes ☐ No	1 2 3 4 (5)	
		• Schedule Posted (e.g., picture schedule or written schedule. Visible, devl. appropriate for grade).	☐ Yes ☑ No	1 2 3 4 5	
2	Establishing and Teaching Expectations	• Rules: Taught & Reviewed (expectations are taught & reminders are provided frequently...used to prevent misbehavior AND after misbehavior occurs).	☑ Yes ☐ No	1 2 3 4 (5)	
		• Active Supervision (e.g., close proximity to students, moving frequently, not stagnant in one location). Actively attending to student behavior, little socializing).	☑ Yes ☐ No	1 2 3 4 (5)	
3	Engaging Students in Observable Ways	• Opportunities to Respond (OTR) (e.g., uses response cards, white boards, electronic responding with a focus on choral responding).	☑ Yes ☐ No	1 2 3 4 (5)	
		• Direct Instruction (modeling the skill, leading students through content with examples and making connections. Assessing student knowledge to guide continued instruction, not relying on worksheets, e.g., "I do, we do, you do").	☑ Yes ☐ No	1 2 3 4 (5)	
		• Class-Wide Tutoring (students are paired strategically to enhance learning, use of response trials, immediate error correction/feedback, teacher actively supervises tutoring).	☑ Yes ☐ No	1 2 3 (4) 5	
		• Computer Assisted Instruction (use of technology that allows each student one-on-one **instruction** without leaving the classroom)	☐ Yes ☑ No	1 2 3 4 5	
		• Guided Notes (lecture or chapter outlines providing main ideas where students can follow along filling in additional ideas, devl appropriate for grade).	☐ Yes ☑ No	1 2 3 4 5	

4	Recognizing Appropriate Behavior using Various Strategies	• **Using Behavior-Specific Praise** (verbal praise clearly identifying student behaviors that earn teacher approval, e.g., "Great job lining up quickly and quietly!"). • **Token Economies** (students earn points or coupons for appropriate/expected behavior that can later be exchanged for prizes or activities). • **Class-Wide Group Contingencies** (a common expectation is set for entire class and all students earn positive outcome if expectation is met, e.g., if everyone works quietly during small groups the class receives extra recess). • **Behavior Contracts** (document written by teacher and student that outlines expected behavior and outcomes when those expectations are or are not met, is devl appropriate for grade, e.g., home-school note, CICO).	☑ ☐ ☐ ☑ ☐ ☑ ☐ ☑	1 2 3 4 (5) 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5					
5	Responding to Inappropriate Behavior using Various Strategies	• **Brief Instructional Corrections for Inappropriate Behavior** (undesired behavior is clearly and immediately identified & an alternative/appropriate behavior is concisely identified, e.g., you pushed Sam, you need to say "excuse me" when you want to pass by). • **Performance Feedback** (students view their performance – compared to a set criterion - visually, e.g., charts, graphs, behavior card. Academic (ORF) or behavioral (e.g., transitions under 2 min, homework turned in). • **Planned ignoring** (identify minor misbehavior to ignore and withhold attention, corrective comment, or redirection when students engage in these minor misbehaviors). • **Differential Reinforcement** (Responding to i.e., commenting, describing, praising appropriate behavior, while ignoring minor misbehavior. Responding to any behavior OTHER than the identified misbehavior). • **Response Cost** (stimulus, e.g., token or coin, is removed due to engagement in undesired behavior) • **Time Out from Reinforcement** (removal from reinforcement, e.g., playing with peers, peer or teacher attention, activity, for a **BRIEF time**, i.e., 10 min or less).	☑ ☐ ☐ ☑ ☐ ☑ ☑ ☐ ☐ ☑ ☑ ☐	1 2 3 4 (5) 1 2 3 4 5 1 2 3 4 5 1 2 3 (4) 5 1 2 3 4 5 1 2 3 4 (5)					

Praise Frequency		Reprimand Frequency			
Behavior Specific	General	Mild	Medium	Harsh	Gesture
𝍸 I	𝍸 III	𝍸 I	I		I

PRAISE DEFINITIONS:	
Behavior Specific	Any **specific verbalization or gesture** that expresses a favorable judgment on an activity, product, or attribute of the student (examples below). *Thank you for sitting criss cross *Good job cleaning up *Nice work helping *That is a pretty picture! *I like how you raised your hand
General	Any **nonspecific verbalization or gesture** that expresses a favorable judgment on an activity, product, or attribute of the student (examples below) *Great! *Nice Work *Hi-five or Thumbs up (gesture w/ no verbalization) *Gives token (no verbalization) *Thank you *Perfect

REPRIMAND DEFINITIONS:	
Mild Reprimand:	Any **verbal comment (using a normal speaking tone)** that indicates disapproval of a student(s) behavior. The verbal comment can be an instruction following student misbehavior. The reprimand is concise (brief). Also referred to as a "redirection" of student behavior. Disagreeing with a student with the **absence of sarcasm or a critical tone** would be identified as mild. *No thank you *Not now *No, come sit down (child at desk, while other children are on carpet) *That is not how we treat our friends
Medium (Sarcastic) Reprimand:	Any **verbal comment (using a sarcastic or critical tone)** that indicates disapproval of a student(s) behavior. The verbal comment can be in the form of a question that is disapproving and has a **mocking, rude, or critical tone**. A sarcastic reprimand is marked if the teacher disagrees with the child using a critical tone. *I don't remember telling you to write about mumkins! *No, it's not cold in here (critical disagreement) *Is that your best work? (critical) (child misspelled word, sarcasm)
Harsh Reprimand:	Any **verbal comment (using a louder than typical tone for the setting)** that indicates disapproval of a student(s) behavior. Harsh reprimand is also marked if the reprimand implies negative consequences (e.g., a threat). *One more outburst and no recess (threat) *I won't tell you again *Excuse me! (loud) *How many times do I need to say _____
Gesture Reprimand:	Any **gesture (without speaking)** that indicates disapproval of a student behavior (e.g., hands on hips). Teacher may also gesture by **physically guiding the child's body** to a preferred area or activity *Hands on hips (disapproving look) *Teacher physically guides child to correct location *Shakes head when student interrupts

Epstein et al., 2008; Office of Special Education Programs, 2016; Simonsen et al., 2008

Appendix D: FCOT

Observer Code: ________________________ Job Title: ________________ Reason for the Observation: _________________________(e.g., re-eval, consultation).

Reliability Observer Code: _________________ Job Title: ___________________ Date: _____________ School Code: ______________ Observation Length:_______

Teacher Code: ____________ Class Description________________________ (e.g., 5th grade general ed). How many students in the classroom _______ ?

Directions 1) **STRATEGY**: Indicate if the classroom strategy was observed during the 20-min observation.

2) **QUALITY**: If a strategy was observed, indicate the quality of the observed strategy (1 = poor, 5 = great).

3) **FREQUENCY**: Using the operational definitions, tally the frequency of Praise and Reprimands observed during the 20-min observation.

	Critical Feature	Classroom Strategy & Description	Observed Yes or No	Quality (Inconsistent with Description … Consistent with Description)	Comments/Notes
1	Maximizing Structure & Predictability	• Easy Traffic Flow (physical arrangement of the room (e.g., adult can easily move in the isle & btw furniture, can access all of room, no roadblocks, no visual obstructions).	☐ ☐	1 2 3 4 5	
		• Structured (teacher-imposed organization, lesson plan is clear, students are aware of current classroom objectives; visual aids are informative and not distracting)	☐ ☐	1 2 3 4 5	
		• Rules Posted (classroom rules are visible, positively stated, large enough to read, succinct, linked to SW expectations).	☐ ☐	1 2 3 4 5	
		• Schedule Posted (e.g., picture schedule or written schedule. Visible, devl. appropriate for grade).	☐ ☐	1 2 3 4 5	
2	Establishing and Teaching Expectations	• Rules: Taught & Reviewed (expectations are taught & reminders are provided frequently...used to prevent misbehavior AND after misbehavior occurs).	☐ ☐	1 2 3 4 5	
		• Active Supervision (e.g., close proximity to students, moving frequently, not stagnant in one location). Actively attending to student behavior, little socializing).	☐ ☐	1 2 3 4 5	
3	Engaging Students in Observable Ways	• Opportunities to Respond (OTR) (e.g., uses response cards, white boards, electronic responding with a focus on choral responding).	☐ ☐	1 2 3 4 5	
		• Direct Instruction (modeling the skill, leading students through content with examples and making connections. Assessing student knowledge to guide continued instruction, not relying on worksheets, e.g., "I do, we do, you do").	☐ ☐	1 2 3 4 5	
		• Class-Wide Tutoring (students are paired strategically to enhance learning, use of response trials, immediate error correction/feedback, teacher actively supervises tutoring).	☐ ☐	1 2 3 4 5	
		• Computer Assisted Instruction (use of technology that allows each student one-on-one <u>**instruction**</u> without leaving the classroom)	☐ ☐	1 2 3 4 5	
		• Guided Notes (lecture or chapter outlines providing main ideas where students can follow along filling in additional ideas, devl appropriate for grade).	☐ ☐	1 2 3 4 5	

4	Recognizing Appropriate Behavior using Various Strategies	• Using Behavior-Specific Praise (verbal praise clearly identifying student behaviors that earn teacher approval, **e.g., "Great job lining up quickly and quietly!"**). • Token Economies (students earn points or coupons for appropriate/expected behavior that can later be exchanged for prizes or activities). • Class-Wide Group Contingencies (a common expectation is set for entire class and all students earn positive outcome if expectation is met, e.g., if everyone works quietly during small groups the class receives extra recess). • Behavior Contracts (document written by teacher and student that outlines expected behavior and outcomes when those expectations are or are not met, is devl appropriate for grade, e.g., home-school note, CICO).	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5					
5	Responding to Inappropriate Behavior using Various Strategies	• Brief Instructional Corrections for Inappropriate Behavior (undesired behavior is clearly and immediately identified & an alternative/appropriate behavior is concisely identified, e.g., you pushed Sam, you need to **say "excuse me" when you want to pass by**). • Performance Feedback (students view their performance – compared to a set criterion - visually, e.g., charts, graphs, behavior card. Academic (ORF) or behavioral (e.g., transitions under 2 min, homework turned in). • Planned ignoring (identify minor misbehavior to ignore and withhold attention, corrective comment, or redirection when students engage in those minor misbehaviors). • Differential Reinforcement (Responding to i.e., commenting, describing, praising appropriate behavior, while ignoring minor misbehavior. Responding to any behavior OTHER than the identified misbehavior). • Response Cost (stimulus, e.g., token or coin, is removed due to engagement in undesired behavior) • Time Out from Reinforcement (removal from reinforcement, e.g., playing with peers, peer or teacher attention, activity, for a **BRIEF time**, i.e., 10 min or less).	☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐	1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5					

Praise Frequency		**Reprimand Frequency**			
Behavior Specific	**General**	**Mild**	**Medium**	**Harsh**	**Gesture**

PRAISE DEFINITIONS:	
Behavior Specific	Any **specific verbalization or gesture** that expresses a favorable judgment on an activity, product, or attribute of the student (examples below). *Thank you for sitting criss cross *Good job cleaning up *Nice work helping *That is a pretty picture! *I like how you raised your hand
General	Any **nonspecific verbalization or gesture** that expresses a favorable judgment on an activity, product, or attribute of the student (examples below) *Great! *Nice Work *Hi-five or Thumbs up (gesture w/ no verbalization) *Gives token (no verbalization) *Thank you *Perfect

REPRIMAND DEFINITIONS:	
Mild Reprimand:	Any **verbal comment (using a normal speaking tone)** that indicates disapproval of a student(s) behavior. The verbal comment can be an instruction following student misbehavior. The reprimand is concise (brief). Also referred to as a "redirection" of student behavior. Disagreeing with a student with the **absence of sarcasm or a critical tone** would be identified as mild. *No thank you *Not now *No, come sit down (child at desk, while other children are on carpet) *That is not how we treat our friends
Medium (Sarcastic) Reprimand:	Any **verbal comment (using a sarcastic or critical tone)** that indicates disapproval of a student(s) behavior. The verbal comment can be in the form of a question that is disapproving and has a **mocking, rude, or critical tone**. A sarcastic reprimand is marked if the teacher disagrees with the child using a critical tone. *I don't remember telling you to write about mumkins! *No, it's not cold in here (critical disagreement) *Is that your best work? (critical) (child misspelled word, sarcasm)
Harsh Reprimand:	Any **verbal comment (using a louder than typical tone for the setting)** that indicates disapproval of a student(s) behavior. Harsh reprimand is also marked if the reprimand implies negative consequences (e.g., a threat). *One more outburst and no recess (threat) *I won't tell you again *Excuse me! (loud) *How many times do I need to say _____
Gesture Reprimand:	Any **gesture (without speaking)** that indicates disapproval of a student behavior (e.g., hands on hips). Teacher may also gesture by **physically guiding the child's body** to a preferred area or activity *Hands on hips (disapproving look) *Teacher physically guides child to correct location *Shakes head when student interrupts

Epstein et al., 2008; Office of Special Education Programs, 2016; Simonsen et al., 2008

Appendix E: Informed Consent

CONSENT TO PARTICIPATE IN RESEARCH –Primary Observer Form
Assessing the Reliability of the Five in 20 Classroom Observation Tool

You are invited to participate in a research study conducted by Danielle Buechlein, School Psychology graduate student, supervised by Margaret Floress, Ph.D. Your participation in this study is entirely voluntary. Please ask questions about anything you do not understand.

Purpose of the Study
The aim of the current study is to assess the reliability of the Five in 20 Classroom Observation Tool (FCOT). The tool was developed in Dr. Floress' lab (the primary investigator's mentor), based on the five critical features and 20 evidence-based strategies identified by Simonsen and colleagues (2008). This tool may prove useful to school psychologists who consult with teachers regarding effective classroom management practices.

Procedures
Participants will view a 20-minute video of class-wide instruction while completing the observation tool. You will further complete a demographics form. You will receive a $10 gift card for your participation.

Potential Risks and Discomforts
This study has been approved by the Eastern Illinois University Institutional Review Board (#). There are no foreseeable risks associated with participating in this study.

Confidentiality
All participant forms will be coded (e.g., A-1) to keep participant data confidential. Your name (or other personal information) will not be paired with your demographic or observation data. Collected data will be emailed to Danielle Buechlein and downloaded onto a password protected computer. All participant data will be stored for at least 3-years. Danielle Buechlein, Dr. Floress, and Jess White (a graduate school psychology, research assistant) will be the only persons with access to data.

Anticipated results are expected to provide information regarding the reliability of the observation tool. We hope that the results from this study will help develop an efficient observation too that school psychologists can use to guide meaningful consultation recommendations.

If you have questions or concerns about this research, please contact: Danielle Buechlein at dabuechlein@eiu.edu or 812-827-1261. If you have any questions or concerns about the treatment of human participants in this study, you may call or write:

Institutional Review Board
Eastern Illinois University600 Lincoln Ave.
Charleston, IL 61920
Telephone: (217) 581-8576
E-mail: eiuirb@www.eiu.edu

I voluntarily agree to participate in this study. I understand that I am free to withdraw my consent and discontinue my participation at any time without consequences of any kind or loss of benefits or services. I have been given a copy of this form.

Participant's Signature Date

Investigator's Signature Date