Exploring Profiles of Coaches' Fidelity to Double Check's Motivational Interviewing-Embedded Coaching: Outcomes Associated with Fidelity

Elise T. Pas, Ph.D.[1]

Lindsay Borden, Ph.D.[2]

Katrina J. Debnam, Ph.D., MPH[3]

Danielle De Lucia, M.P.S[1]

Catherine P. Bradshaw, Ph.D., M.Ed.[4]

[1]Johns Hopkins University, Bloomberg School of Public Health

[2]Johns Hopkins University, School of Medicine

[3]University of Virginia, School of Nursing

[3]University of Virginia, School of Education and Human Development

Published: June 2022

ACKNOWLEDGEMENT: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150221 (PI: C. Bradshaw) to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors would like to thank the coaches for this study, Ms. Sandy Rouiller, Dr. Lana Bates, Dr. Dana Darney, Dr. Lauren Kaiser, Ms. Brenda Kelly, Mr. Jeff Krick, Dr. Kristine Larson, and Ms. Rebecca Piermattei. We also thank Drs. Stephanie Moore, Rashelle Musci, Joseph Kush for consultation on the statistical approach.

Abstract

Motivational interviewing (MI) is applied in a variety of clinical and coaching models to promote behavior change, with increasing interest in its potential to optimize school-based implementation fidelity. Yet there has been less consideration of fidelity indicators for MI-embedded coaching and links to outcomes. We leveraged secondary data from 151 teachers across 18 schools, which were part of a larger 39 middle school randomized controlled trial of a teacher coaching model, to explore profiles of fidelity and the associations between fidelity and outcomes. We conducted latent profile analysis (LPA) to examine profiles of four components of fidelity (i.e., adherence, dosage, quality, and teachers' responsiveness). Next, we examined whether observed teacher practices and student behaviors varied across fidelity profiles. Because coaches and independent coders reported adherence, we also examined the reliability of retrospective coach adherence ratings. Results indicated that coaches show promise as a reliable rater of adherence. We identify concrete areas to ensure that reliability can be achieved in other contexts. The LPA indicated that there were two (high and lower) fidelity profiles. Statistically significantly fewer instances of student non-cooperation were observed in classrooms where the teacher was engaged in high fidelity coaching, reflecting a large effect size. Moderate-sized, but non-statistically significant, effects also emerged for teacher opportunities to respond and reactive behavior management. Future directions are considered regarding fidelity measurement and how to optimize coaching.

KEYWORDS: adherence, dosage, quality, responsiveness, teacher practices, randomized controlled trial

Exploring Profiles of Coaches' Fidelity to Double Check's Motivational Interviewing-Embedded Coaching: Outcomes Associated with Fidelity

There is increasing recognition of the potential utility of Motivational Interviewing (MI) in a variety of clinical interventions, with even more recent interest in its application in school-based coaching (Pas & Bradshaw, 2021). Specifically, MI is used in coaching models to enhance teachers' feelings of efficacy to change their practices and help empower them to overcome ambivalence (Miller & Rollnick, 2012). MI utilizes collaborative and non-judgmental language supportive of change to promote participants' change talk (i.e., language in support of new behaviors). Given its flexibility and applicability, MI has been studied in a range of contexts, with varying audiences, to address numerous outcomes (Frey et al., 2021). Yet only recently has there been attention to the fidelity of MI embedded within coaching. Research has not yet explored variations among different raters of fidelity, such as comparing coach self-report and external observations of the coaching process, and there has been limited study of how coaching fidelity indicators are related to outcomes. Instead, extant coaching literature that has focused on fidelity has examined whether coaching impacts teacher fidelity of implementation of evidence-based practices (e.g., Cross et al., 2015; Stahmer et al., 2015; Sutherland et al., 2015, 2018).

However, there is a paucity of fidelity research, and little has focused on a multi-dimensional measurement of fidelity. For example, a recent literature review suggests that a mere 10.7% of adult and child therapy studies published included data on fidelity (Cox et al., 2019). Yet this marked a nearly threefold increase from a prior study in this field, conducted 12 years earlier (Perepletchikova et al., 2007), suggesting a positive trend toward recognition of the importance of fidelity. Moreover, a review focused on teacher coaching to support the implementation of social and behavioral interventions indicated that 31% of reviewed studies

included fidelity measurement (Stormont et al., 2015). Although the school-based coaching and MI research has a stronger emphasis on fidelity measurement (see Magill et al., 2014, 2018), relative to clinical research, fidelity remains a major gap in educational research on coaching. More consistent examination of coaching fidelity, the integration of multiple indicators of fidelity, and the inclusion of quality indicators (e.g., quality of MI when embedded into coaching) are needed in the school-based coaching literature.

The current study sought to address these gaps by investigating multiple indicators of four fidelity components for an MI-embedded school coaching model. Specifically, we examined coach self-report, independent coders' observations, and teacher reports regarding coaching adherence, dosage, quality, and participant responsiveness (see definitions below; Dane & Schneider 1998). We first compared coach self-reported adherence and independent coder ratings of adherence to inform the field about reliability of coach self-reports. We then examined patterns or "profiles" of fidelity of the coaching, using a person-centered analytic approach (i.e., latent profile analysis) to identify clusters of teachers who received similar levels of coaching fidelity across measures. Finally, we explored whether there were mean differences in observed teacher practices and student behaviors across these latent profiles of fidelity. This line of research has the potential to inform research on the fidelity of coaching broadly and of MI-embedded coaching to increase program implementation and student outcomes.

Use of Motivational Interviewing in Coaching

The goal of teacher coaching is to support teacher development of new practices and skills (Joyce & Showers, 1980). MI has been embedded into some coaching models as an implementation strategy "to enhance the adoption, implementation, and sustainment of a new…practice" (Lyon et al., 2019, p. 66). Although MI was first developed and tested in the

addiction counseling field, the focus of MI on addressing ambivalence to change is of great relevance in schools and other contexts. Not surprisingly, MI has gained considerable traction in schools in recent years (Herman et al., 2020).

A core tenet of MI is that ambivalence toward change is normative (Miller & Rollnick, 2012). A recent article demonstrated this to be true for teachers during coaching, where teacher volleying between change and sustain talk (i.e., language in support of the status quo) was observed (Pas et al., 2021). Although research on MI fidelity is extensive in the substance use literature (e.g., see review by Magill et al., 2018), research on MI in schools has been limited. A recent special issue on MI in *Prevention Science* included a few school-based MI studies that contributed to the field by identifying how MI can be used within a group-based approach as an implementation strategy (Larson et al., 2021); how the measurement of teacher confidence and importance rulers (an MI strategy) related to subsequent teacher practices (Owens et al., 2021); and how coach use of MI related to teacher change talk (Pas et al., 2021; for a review of the special issue, see Pas & Bradshaw, 2021). Despite these important contributions, there remain a range of questions regarding fidelity to coaching and MI-embedded coaching, as well as how fidelity relates to teacher and student outcomes.

Key Definitions for MI Fidelity Research

Fidelity, defined as the degree to which the intervention is being practiced as intended, has been well established in the literature (Dane & Schneider, 1998). Despite this, as noted earlier, the measurement of fidelity remains infrequent and incomplete and has been noted as vague to define and measure (Perepletchikova et al., 2007; Schoenwald et al., 2011). Additionally, extant research (Durlak & DuPre, 2008; Stormont et al. 2015) suggests that fidelity is linked with better outcomes. In considering fidelity of MI-embedded coaching, we leverage

Dane and Schneider's (1998) model of the five related but distinct core components of fidelity measurement. Specifically, *adherence* refers to how closely the implementer follows program guidelines while also avoiding contraindicated procedures. This can be measured in the form of a checklist, with discrete activities identified that a rater responds to the presence or absence of core components. *Dosage* is defined as the amount of the intervention delivered by the implementer. In the case of coaching, this could be conceptualized as the amount of time spent coaching. *Quality of implementation* reflects the degree to which program delivery impacts the identified program goals (e.g., teacher behaviors, student outcomes) and represents a more challenging construct to both define and measure. *Participant responsiveness* includes the level of engagement and interactivity implementers elicit from their participants. In the case of MI-embedded coaching, this would be the teacher engagement within the coaching process and could be measured in a range of ways (e.g., teacher report, coach report, coding of language within a session). Finally, *program differentiation* occurs when key program components are present and serve to distinguish the program from other practices or procedures. For example, in teacher coaching, this may include the measurement of all other professional development activities that both intervention and control teachers receive to determine whether the addition of coaching reflects a distinct additional exposure as compared to the control group. As can be seen, following the conceptualization of these fidelity components can assist with the identified concerns in defining fidelity in a more concrete way (i.e., as identified in Perepletchikova et al., 2009) and have direct impacts on a measurement plan.

Gaps in Fidelity Measurement and Research

For each of these five components, measurement strategies can vary widely based on program context. In a survey of 90 corresponding authors for psychotherapy studies, the barriers

of collecting fidelity data included lack of resources, issues in designing measures and analyzing data, and challenges in training clinicians to complete them (Perepletchikova et al., 2009). Similarly, naturalistic settings, such as community agencies and schools, often struggle with limited financial resources and shortages of skilled personnel. As such, any implementation fidelity assessment strategy that requires substantial time or financial investment (e.g., observations, coding) is not as viable and has led to an over-reliance on self-report or no fidelity measurement. Such complications in the delivery and completion of fidelity assessments are notable within coaching or consultation research (Sheridan et al., 2009) because, by nature, coaching is tailored to individuals and is generally delivered within complex and "real-world" settings (e.g., schools). However, many evidence-based coaching and consultation models follow a staged problem-solving model including rapport building, identifying a problem, and intervention selection, implementation, and evaluation (Erchul & Sheridan, 2008; Solomon et al., 2012) and fidelity to each part can be measured.

Fidelity measurement strategies can be either direct or indirect (Schoenwald et al., 2011). Direct measures are often observational, requiring audio/video recordings or in-vivo observations by trained observers. Indirect measures may include questionnaires or checklists completed by the implementer (e.g., coach), participant (e.g., teacher), or trained coders through an external research process. The direct approach may be best suited for measuring micro-level fidelity components, such as specific verbal or non-verbal behaviors, while the indirect methodology can measure presence or absence of intervention components within a session (Heaton et al., 1995). Direct measures of fidelity are expensive and time consuming (Bishop et al. 2014); indirect measures also face barriers, including being subject to bias (i.e., self-reports) and burdensome for participants to complete. A recent literature review of coaching studies

found that just 9 out of 29 studies reviewed included direct measures of coaching fidelity (Stormont et al., 2015). Typically, an independent observer collected data using a scripted checklist either in person or by listening to audiotapes for a certain percentage or number of observations; studies utilized these data to report high levels of coach fidelity (95%–100%; Stormont et al., 2015).

Similarly, in reviews of MI studies focused on students, one review found that adherence checklists were collected in 4 out 11 studies and observations were conducted in 2 out of 11 studies (Snape & Atkinson, 2016) and the other indicated that just 7 out of 20 original research studies included fidelity coding, mostly using established MI measures (Mutschler et al., 2018). In another review of 14 studies, seven included observations of fidelity (i.e., direct), five relied on teacher self-report (i.e., indirect), and two studies used both (Dusenbury et al., 2003). In psychotherapy research, indirect methods are rarely correlated with each other when aiming to measure a single construct (Heaton et al., 1995; Waltz et al., 1993), which could arise from issues of social desirability bias (Donaldson & Grant-Vallone, 2002) or because data forms are not completed rapidly enough (e.g., they are completed by memory days or even months later; Bishop et al., 2014). It is possible that well-trained and supervised coaches providing data immediately following a session may be reliable in self-reporting their adherence to a program, but this remains an area for further research.

As noted previously, quality is perhaps the most challenging fidelity component to measure. Regarding MI specifically, there are a variety of measures of MI used within substance use research but not in school-based models. Given the dynamic process of MI between a clinician/implementer and implementer, a measure excluding participant responsiveness is missing a key aspect of MI. In terms of coding MI, the literature is mixed regarding the utility of

measuring MI-consistent language globally (i.e., collapsing across skills) or measuring distinct MI skills (e.g., OARS; O = *open-ended questions*, A = *affirmations*, R = *reflections*, and S = *summaries*). Studies of collapsed "MI-consistent" langauge have documented significant associations with change talk with greater consistency than studies examining the OARS discretely (e.g., Magill et al., 2014, 2018; Moyers & Martin, 2006). Studies examining discrete MI skills have greater nuance and potential for practice implications. Unfortunately, the findings of such studies varies, where some (e.g., Apodaca et al., 2016) have demonstrated that all skills are associated with change talk, but others indicate that the quality of reflections (i.e., simple versus complex) differentially relate to positive outcomes (see Laws et al., 2018). Furthermore, the tradeoff for this greater nuance are increased challenges in rater reliability and transportability (see Pas et al., 2021). See Frey et al. (2021) for a full review of measures.

A final key consideration for fidelity research is the calculation of fidelity scores, namely adherence. Cordray and colleagues (e.g., Cordray & Pion, 2006; Hulleman & Cordray, 2009) outlined a few adherence indices for consideration, all of which examine fidelity across a number of items/indicators: (a) *average fidelity index* (i.e., average "points" earned), (b) *absolute fidelity index* (i.e., comparing actual implementation to the absolute, or maximum, level of fidelity), and (c) *binary complier index* (i.e., a yes/no response as to whether implementation reached a specific [hypothesized or empirical] cut point for elements implemented; Hulleman & Cordray, 2009). Much adherence reporting uses the average fidelity index (for a review, see Stormont et al., 2015), likely because of the ease with which can be created and interpreted.

The Present Study

The present study aimed to (1) examine the reliability of coach self-reports for adherence as well as highlight distinct areas of disagreement to identify areas in need of attention, (2)

identify profiles of fidelity for coaching utilizing multiple indicators of 4 out of 5 fidelity components (i.e., adherence, dosage, quality, and participant responsiveness), and (3) examine how these fidelity profiles related to teacher and student outcomes. We anticipated that coaches would demonstrate adequate adherence and may be a reliable source of data, given the data came from a randomized controlled trial, where coaches were trained and supervised to complete checklists following their sessions. Based on prior research on Double Check, which examined variability in coaching dosage and in MI usage and change talk (e.g., Pas et al., 2016; Pas et al., 2021), we expected that we would at least see two profiles of fidelity (i.e., high and low) but were specifically interested in determining whether there were more nuanced profiles. For example, given that the coaching relationship and process can vary and be tailored based on teacher need and responsiveness, we were interested to see whether a range of patterns emerged, such as instances of high fidelity across all components, versus instances of some components being high (e.g., adherence) whereas others (e.g., quality or responsiveness) were low. Given the expected and prior reported impacts of the Classroom Check-Up on both teacher practices and student behaviors (e.g., Bradshaw et al., 2018; Reinke et al., 2008), we expected that higher fidelity would be associated with the most positive outcomes.

Method

Research Design and Procedures

This study utilized secondary data from 18 out of 19 intervention schools participating in the Double Check (Bradshaw et al., 2018) efficacy trial conducted in 2015–2020. One intervention school had just four eligible teachers and none completed data. Across four years (i.e., for the 2015-16 through 2018-19 school years), a point of contact for the district was approached each year to determine interest in participating in the trial and provided district

approval. Districts assisted in recruitment by identifying schools and recruiting principals to consent for their schools' participation in the project. Following a formal overview of the study, consenting principals signed a commitment letter outlining the randomization, intervention details, and data collection protocols. Once a school was recruited, core subject classroom teachers (i.e., English Language Arts, Math, Science, and Social Studies) were provided information about the study and asked to provide formal written consent, if they wanted to participate in the study. Primarily sixth and seventh grade teachers were recruited, however eighth grade teachers were included in some schools. Following baseline data collection, schools were randomized to intervention or control conditions. Consented teachers in schools randomized to the intervention condition were additionally asked to provide written consent for coaching session audio recordings (i.e., they could participate without providing consent for audio recording). The researchers' Institutional Review Board approved this study.

Participants

Teachers in the intervention schools were included in this study because they were assigned to receive coaching. There were 153 intervention teachers, 128 of whom who completed coaching (i.e., three teachers left the school during the study, 10 teachers consented but then declined coaching, four teachers delayed starting and then ran out of time to be coached, and we did not have records on eight teachers' reasons for not completing coaching). We retained 151 teachers with data for these analyses as even most non-coached teachers had baseline survey data and dosage data (i.e., zero minutes); no significant differences emerged between the full and coached samples (see Table 1 for sample descriptives for the full 153 and 128 teachers). Among the full sample of 153 teachers, almost one-third of the teachers self-reported that they had been teaching for 4–8 years, followed by about one-quarter of teachers

who had been teaching for 1–3 years, about one-fifth who were first year teachers, and just under 15% taught for 9 or more years (see Table 1 for exact sample sizes). The largest proportion (under half) of the sample was comprised of White teachers followed by Black/African American teachers (over one-third); much smaller portions reported their race/ethnicity as Asian/Pacific Islander, Hispanic/Latino, and "other" (see Table 1). Sixth, seventh, and multiple grade teachers comprised the largest and close to equal proportions of the sample (see Table 1). Similarly, relatively consistent portions (i.e., one-quarter to one-third) of teachers taught Math, Science, English Language Arts, and Social Studies (listed in descending order of proportions). Regarding age, one-quarter each reported being 20–30 and 31–40, just under one-fifth reported being 41–50 years old, and under one-fifth reported being 51 or older. Teachers were relatively evenly spread across the 4 consecutive study cohorts, reflective of cohort and school size differences. Most teachers (i.e., over 90%) were female. See Table 1.

Eight coaches provided support to the four cohorts of teachers; six coaches worked with just one cohort and two worked with two cohorts. Seven coaches were female; two coaches were Black/African American and six were White. All coaches had prior coaching experience. Their educational backgrounds were either in education (i.e., four B.A. or M.A.; one Ph.D.) or they had a Ph.D. in School Psychology (i.e., three). Coaches were supervised bi-weekly by the first author or lead coach, who listened to session audio and provided structured feedback.

Intervention: Coaching Component of Double Check

The Double Check framework uses coaching, school-wide professional development sessions, and support for school-wide positive behavior supports to improve teacher implementation of culturally responsive practices and increase student engagement. However, the focus of the present study was on the coaching component of the Double Check (see

Bradshaw et al., 2018). Double Check coaching was an adaptation of the Classroom Check-Up (CCU) model (Reinke et al., 2011), in which MI was embedded to promote teachers' feelings of efficacy to change their practices and help empower them to overcome ambivalence (Miller & Rollnick, 2012). MI was explicitly built into the staged problem-solving to promote fidelity to its use. In the interview, the coach focused on rapport building, engaging the teacher, and evoking desired classroom changes (Miller & Rollnick, 2012; Reinke et al., 2011). For Double Check coaching, data collection consisted of the coach visiting each classroom three times, where they tallied teacher and student behaviors. Additionally, the teacher provided a self-reported checklist about the classroom. Key areas of data collection included general positive behavior supports and classroom management (e.g., setting clear expectations, providing behavior-specific praise), instructional management (e.g., opportunities to respond), and culturally responsive practices including **C**onnection to the curriculum, **A**uthentic relationships, **R**eflection on practices, **E**ffective communication, and **S**ensitivity to students' culture (i.e., discussed in the model using the acronym CARES; see Bradshaw et al., 2018).

After data collection, the coach provided integrated feedback for about 30–45 min in a one-on-one session, utilizing the classroom visit and teacher checklist findings. The goal was to help the teachers process the information and select areas to work on; MI strategies were leveraged to evoke teacher reasons to select an area and record a menu of options for intervention. Following feedback, the teacher and coach engaged in goal setting and action planning, which also embedded scripted MI techniques (e.g., importance and confidence rulers) to evoke change talk and commitment language (Amrhein et al., 2003). Teachers then received ongoing performance feedback and check-ins for progress monitoring and goal refinement.

Measures

Measures included those assessing the four fidelity components (i.e., adherence, dosage, quality, and participant responsiveness) by multiple methods and raters (i.e., seven indicators in total), as well as teacher and student outcome measures. Below we summarize each in detail.

Fidelity coding

Adherence. After each interview, feedback, and goal setting/action planning session, coaches completed an implementation adherence checklist indicating whether they had excluded (0), partially implemented (1), or fully implemented (2) each item (see Pas et al., 2016; Pas et al., 2022). For the individual interview, this included 7 items (e.g., asking each question, explaining the coaching process). There were 8 items for the feedback session (e.g., explaining the feedback form, summarizing data, asking for input throughout), and 10 items for the goal setting session (e.g., reviewing the action planning process, prompting the teacher to set two goals, asking the confidence and importance rulers). The percent of all items that were rated as implemented with full fidelity, by the coach and external coder, were used to calculate adherence (i.e., two variables). Using audio tape recordings of these sessions, research assistants were trained to also complete these checklists. Of the 253 coded sessions (i.e., 82 interviews, 87 feedback sessions, and 84 action planning sessions), 60 (23.7%) were double coded: 21 interviews with inter-rater agreement at 94%, 20 feedback sessions with inter-rater agreement at 86%, and 19 feedback sessions with inter-rater agreement at 89%.

Dosage. After each contact with an intervention teacher, the coaches completed an online coaching log, documenting the time spent with each teacher. They also logged specific activities, including time spent on each step of the CCU coaching process (i.e., the interview, training the teacher on the data collection procedures, classroom visits for data collection, feedback, action planning, and follow-up observations and feedback) and time spent in relationship building

activities and coach preparation and planning time. We aggregated the total number of minutes and count of contacts with each teacher, as two indicators of dosage.

Implementation Quality. Audiotaped coaching feedback sessions were used to assess implementation quality, because this was where the most variability in MI use was introduced (see Pas et al., 2021). All audio recordings were transcribed; coding was completed by reviewing the transcriptions and listening to the audio to optimize coders' reliability. Audio recordings were linked to the data entry system (Procoder; Tapp & Walden, 1993) and transcriptions were used to ensure accuracy. The Procoder system timestamped all codes in real time and allowed for pausing and restarting. About one-fifth of the 87 available feedback session recordings were double coded to check reliability; the coders achieved 80% interobserver agreement (i.e., number of agreed-upon codes within 5 s of one another, divided by the total number of utterances coded by either rater; see Pas et al., 2021, for additional details). There was a separate file for each session recording with timestamped utterances and codes that were merged and analyzed.

Coding of coach and teacher language was completed using an adapted and integrated version of a commonly used MI coding system, the MI-SCOPE (Martin et al., n.d.; Moyers & Martin, 2006; see Pas et al., 2021, for details on the adapted measure). The adaptations of the MI-SCOPE included collapsing the 30 MI implementer (coach) codes into (a) MI-consistent, (b) MI-inconsistent, (c) feedback, or (d) other language. The 16 MI-SCOPE participant (teacher) codes were collapsed into e) change talk, f) sustain talk, or g) other language following the system developed by Borden (2012; see Pas et al., 2021, for additional detail). All decipherable (i.e., audible) utterances were coded as one of these seven codes. An utterance was defined as a cohesive thought that either ended because the thought was complete (e.g., and a new thought

began with the same speaker) or because the other speaker interjected an utterance. Facilitative sounds, like “mmm”, “hmm”, or “yeah”, were not coded. All utterances only received one code.

We conceptualized coach use of MI-consistent language as the indicator of quality, which captured the key communication skills within Motivational Interviewing often referred to as as “OARS” (see Miller & Rollnick, 2012; Reinke et al., 2011). Any instance of the following were coded as an MI-consistent utterance: (a) an open-ended question (i.e., to demonstrate curiosity and neutrality), (b) affirmation (e.g., genuine acknowledgement of the teachers strengths, efforts, actions, progress, or values that convey positive regard and caring), (c) reflection of what the teacher said to depict empathic listening and understanding, (d) a summary of the teacher’s statement to check for understanding or transition to a new topic, emphasizing teacher control and autonomy, and (e) permission seeking to provide advice (see Pas et al., 2021). We were particularly interested in the proportion of coach codes that were MI-consistent (i.e., total number of MI-consistent codes for each teacher divided by the total number of coach codes, including MI-inconsistent, feedback, and other language) and this served as the indicator of quality. MI-inconsistent langauge included examples of confrontation, direction, opining, advising without permission, and warning. Feedback included the presentataion of objective data or referencing of information the teacher provided in a prior session. Other language included close-ended questions, sharing of general information or about the structure of the coaching process, self-disclosures, sharing concerns, small talk, etc. (see Pas et al., 2021).

Participant Responsiveness. When using MI, OARS are intended to build rapport, support autonomy, and ultimately to guide the conversation to evoke change talk (Miller & Rollnick, 2012), which is defined and measured as langauge that indicates movement toward a new positive behavior. Since a main goal of MI is to faciltiate change talk, we conceptualized

this as participant responsiveness or an indicator of the teacher's positive engagement. Therefore, we also used the adapted MI-SCOPE (discussed above) for this dimension. All language that included examples of teacher commitment (e.g., "I will" statements), desire (e.g., "I wish/want" statements), expressions of ability (e.g., "I can" statements), need, reasons for and benefits of change, or actual actions toward change were coded as change talk. We were interested in the proportion of this change talk langauge relative to all other teacher language, including sustain talk (i.e., indications of wanting to maintain the status quo) or any other language (i.e., coded similarly to coach other language; see Pas et al., 2021) as one indicator of responsiveness.

A teacher self-report scale on the working relationship with the coach served as an additional indicator of participant responsiveness (Johnson et al., 2016). We selected this specific alliance scale as it was focused on key targets of MI including perceptions of the collaborative relationship established. This scale included 6 items focused on trust, coach approachability, and the coach's understanding of the teacher's goals and views ($\alpha = 0.93$). Example items from this scale include "The coach and I agreed on what the most important goals for intervention were" and "The coach and I trust one another". Teachers rated items on a scale of 0 (*never occurred*) to 4 (*always occurred*). Prior research indicated there was the greatest cross-informant correlations for this scale as well as acceptable consistency and reliability between teacher and coach raters (Johnson et al., 2016). Research focused on the similar construct of therapeutic alliance demonstrates that establishing a collaborative working relationship is associated with client behavior change (e.g., Ackerman & Hilsenroth, 2003) and therefore is important to examine within school-based coaching (see Johnson et al., 2016).

Outcomes

All outcomes were collected through observations conducted by observers hired by the research team. Observers were unaware of the purpose of the study and of school assignment to the intervention status. This included the Classroom Assessment Scoring System –Secondary version (CLASS-S; Pianta et al., 2008) and the Assessing School Settings: Interactions of Students and Teachers (ASSIST; Rusby et al., 2001, 2011). CLASS trainers at Teachstone provided the CLASS-S training to hired observers (including didactic learning and practice opportunities). Observers then had three attempts to meet 80% reliability standards for video recordings, following CLASS certification procedures. Training for the ASSIST was conducted by the research team and involved didactic sessions to review the manual (i.e., including classroom data collection procedures and detailed information on the codes) and coding videos for practice. Observers needed 80% agreement with the expert coder on three live in-school observations to complete training. Video recalibration was conducted during study observation completion and averaged 87.25% for the baseline time point and 83.6% for post-test.

Each participating teacher was observed on three occasions: (a) at baseline (i.e., in the fall of their participating school year), prior to the intervention implementation, and (b) at post-test (i.e., during the spring and at the end of school year). These three observations occurred on different dates and times. Observations took about 60 min in total. Observers began with the CLASS-S, which included a 15-min timed observation for observers to take notes about the classroom which were used to then record scores for 12 dimensions (i.e., spanning another 10–15 min). Dimension scores were averaged to create three composite scores (see below). Observers then completed the ASSIST, which included a 3-min acclimation period and allowed for recording classroom demographics and content information, followed by a 15-min timed session to tally behaviors. The observer then left the classroom to complete global ratings. For each time

point, the scores (i.e., CLASS dimension scores and ASSIST tallies and global ratings) from the three observations were averaged into one single score.

Teacher Practices. As noted above, the CLASS-S is comprised of 12 dimension scores including Positive Climate, Teacher Sensitivity, and Regard for Adolescent Perspectives (which are averaged for the *Emotional Support* composite); Negative Climate, Behavior Management, and Productivity Perspectives (which are averaged for the *Classroom Organization* composite); Instructional Learning Formats, Content Understanding, Analysis and Problem-Solving, Quality of Feedback, and Instructional Dialogue dimensions (which are averaged for the *Instructional Support* composite); and Student Engagement, which is not a part of any composite. Each dimension was rated by the observer on a 1–7 scale (i.e., 1 or 2 indicating *low* quality; 3–5 indicating *middle* quality, and 6 or 7 indicating *high* quality; Pianta et al., 2008).

The three CLASS-S composite scores were of interest to this study. The *Emotional Support* composite captures key relational processes (e.g., positive tone) and captures how well teachers respond to students; this composite relates closely to the targeted Double Check domains (e.g., Authentic relationships and Sensitivity to student's culture). The *Classroom Organization* composite captures classroom management (e.g., time management, keeping student attention and focus) and aligns to the original CCU elements addressing positive behavior supports and classroom management. The *Instructional Support* composite captures instructional structures and approaches, including feedback provided, intended to optimize student engagement for deep learning. This composite is somewhat addressed within the coaching focus on opportunities to respond/pacing and through CARES on the connection to curriculum aspect. Prior CLASS-S research has demonstrated that CLASS-S composite reliabilities (i.e., intraclass correlations) range from good (.73 for Instructional Support) to

excellent (.77 for Emotional Support, .82 for Classroom Organization; Allen et al., 2013). Similarly, ICCs for the dimensions have been reported as good to excellent (.64–.78; Allen et al., 2013). In the current study, ICCs at baseline for the three composites ranged from .60 to .63. See Table 2 for baseline data.

We were also interested in external ratings of tallied teacher practices that were among practices the teachers may have worked on with the coach. For these, we utilized the Assessing School Settings: Interactions of Students and Teachers (ASSIST; Rusby et al., 2001) that includes both event-based tallies (i.e., behavior counts) and global ratings on items that load onto scale scores (see Rusby et al., 2011). The tallies of interest here included (a) proactive behavioral management, (b) opportunities to respond, (c) approval, (d) disapproval, and (e) reactive behavior management. *Proactive behavioral management* included demonstrations of behavioral expectations prior to a problem behavior emerging and included verbal (e.g., explaining, reminding, commanding, prompting) and physical (e.g., modeling) instances. *Opportunities to respond* were any behavioral or instructional prompt to students that required immediate response either spoken verbally or through a publicly shared response (e.g., on a white board); it did not include privately written responses (e.g., in a notebook). *Approvals* were instances of the teacher recognizing students through a tangible item, verbal praise, approving gestures (e.g., thumbs up), or physical contact (e.g., pat on the back). *Disapprovals* displayed dissatisfaction with behavior through a threat or use of a punitive consequence (e.g., detention), verbal criticism/sarcasm, or gestural or physical contact. *Reactive behavior management* included redirections of inappropriate student behavior through a range of cues including touch, gesture, proximity, or commenting, but did not include disapprovals.

Global ratings were collected for a range of teacher practices on the ASSIST as well, but the two of interest here were the *culturally responsive practices* and *positive behavioral supports* scales, as they were unique from any data collected on the CLASS-S or ASSIST tallies and were directly related to coaching foci. The *culturally responsive practices* scale is a 7-item scale and assessed specific practices such as "Teacher connects lessons to real world examples", "Teacher engages in storytelling or sharing", and "Teacher employs rhythm or 'call and response' instructional strategies." Response options for these items were on a 0–4 Likert-type scale (i.e., *never* to *almost continuously*). The *positive behavioral supports* (PBS) scale examined the presence of seven PBS features on a scale of 0 = *no evidence*, 1 = *partially in place*, and 2 = *fully in place*. Examples of features assessed include "3–5 positively stated behavioral expectations (e.g., rules, code of conduct) are posted in the classroom" and "Observed evidence that the teacher has a reinforcement system to reward positive behaviors." Higher scale scores reflect observed higher use of culturally responsive teaching strategies and positive behavior supports. Baseline data are presented in Table 2.

Student Behaviors. A couple of tallies on the ASSIST addressed student behaviors and were of interest here. Specifically, tallied instances of *student non-cooperation* (i.e., a teacher made a request that a student did not respond to within 5 s) and *student disruptive behaviors* (i.e., any initiation or extension of a behavior that interfered with another student, students, the classroom, or the teacher) were tallied. The total tallied instances of these two student behavioral difficulties were outcomes of interest (see Table 2).

Data Analysis

Reliability and Descriptive Analyses

To examine the reliability of coach ratings of adherence, we calculated inter-observer agreement for the checklists for each session and across all sessions. We were unable to calculate Kappa reliability estimates for this measure given that there was a narrow range of possible values (i.e., 0, 1, or 2) and there was restriction of range and a ceiling effect, reflecting high levels of adherence and the "prevalence problem" (Hallgren, 2012, p. 28). Descriptive statistics allowed us to examine specific areas of discrepancy between raters and areas of lower adherence.

Latent Profile Analysis

We conducted latent profile analysis (LPA; Hagenaars & McCutcheon, 2002; Muthén & Muthén, 1997-2012) in M*plus* 7 to examine the profiles of fidelity, using seven indicators of four components (i.e., two indicators of adherence: coach and independent ratings; two indicators of dosage: the total minutes spent in coaching and the number of coach contacts; one indicator of quality: the percent of coach utterances that were coded as MI consistent; and two indicators of participant responsiveness: the percent of teacher utterances that were coded as change talk and teacher self-reports of the working relationship). LPA creates latent or non-observable profiles based on similarities in continuous response scores. In these analyses, we allowed means and variances to vary across profiles and had indicator variances freely estimated. Although there were few statistically significant correlations among fidelity indicators (i.e., of the 21 correlations, only five were statistically significant), we examined nested models whereby none of the fidelity variables were covaried as compared to a model where (1) statistically significant correlations were modeled as covarying using the WITH statement, (2) within measure variables were specified as covarying and teacher responsiveness variables were covaried with all other fidelity variables, and (3) all variables within the LPA were covaried. These three nested models

were compared to the fixed model (i.e., without WITH statements) utilizing the Satorra-Bentler scaled chi-square difference test (Satorra & Bentler, 2010).

We standardized all variables so they could be analyzed and interpreted using the same metric. The LPA analyses were iterative, where we added one profile at a time and considered multiple fit indices and statistical tests to determine the best-fitting model (i.e., number of profiles). We accounted for clustering of teachers and classrooms within schools using the Huber-White sandwich estimator to adjust the standard errors and utilizing the "type=complex mixture" command (Muthén & Muthén, 1997-2012).

The number of profiles was determined by considering the following: the Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC; Schwartz, 1978), sample size adjusted BIC, Lo-Mendell-Rubin likelihood ratio test (LMR; Lo et al., 2001), and Vuong-likelihood ratio test (VLMR; Muthén & Muthén, 1997-2012). Improved fit is indicated by a decreasing AIC, BIC, and adjusted BIC, as well as a statistically significant ($p < 0.05$) LMR and VLMR. The leveling off, or diminishing gain, in BIC and ABIC are considered a criterion for selecting the best-fitting profile model (Masyn, 2013; Nylund et al., 2007). This is further examined by calculating the Bayes Factor (BF) and correct model probability (cmP), where larger BF values (>10; Wasserman, 1997) and cmP values (Masyn, 2013) indicate better fit. The VLMR and LRT tests examine the fit of a given model with a solution with one fewer latent profile; non-significant *p*-values indicate that the specification of an additional latent profile does not result in statistically significant improvement in model fit. Additional model considerations included entropy scores above 0.80 and closest to 1.00 and posterior probabilities greater than 0.70 (Nagin, 2005; Ramaswamy et al., 1993). Finally, we examined the percentage of participants in profiles to ensure that sizable group sizes were included in each profile (Muthén,

2004; Nylund et al., 2007) and determined the distinctive value of and interpretability of the identified profiles (Masyn, 2013). Given our accounting for nesting, the bootstrapped likelihood ratio test (BLRT) was not estimated by M*plus*.

Once the final LPA model was selected, the auxiliary function was used to determine whether there were differences on observed teacher practices (i.e., ASSIST tallies and CLASS-S composite scores) and student behaviors (i.e., ASSIST tallies; Asparouhov & Muthén, 2013). To examine mean differences between profiles on distal outcomes with unequal variances, the DU3STEP function was used. In the output, we examined means and chi-square testing for each outcome. We calculated Cohen's *d* effect sizes (Cohen, 1988) by subtracting the mean of each outcome for the two identified profiles and dividing by the pooled standard deviation. Effect sizes of less than 0.20 are considered *small*, 0.20–0.50 are considered *moderate*, and 0.60 or higher are considered *large* (Cohen, 1988). To ensure that baseline differences on outcomes were not present, in a separate analysis, we used the R3STEP function to examine whether teachers were more likely to be in a specific profile based on their baseline ASSIST and CLASS-S data.

Results

Ratings of Adherence

We calculated the agreement between coach and independent raters for each fidelity checklist item, session, and for all sessions in total when we had both the coach rating in real time and an independent coding completed by listening to an audio recording of the session. Audio recordings were not available for every session (e.g., teacher declined, or the audio did not work). The average agreement for interviews (n = 73) was 77.20% (SD = 15.81). For the feedback session (n = 86), the average agreement was 79.22% (SD = 17.74). For the action planning session (n = 78), the average agreement was 78.79% (SD = 20.28). The average across

all sessions, in instances that reliability was present for at least two sessions (n = 85), was 78.31% (SD = 13.80). On average, coaches rated themselves as fully implementing 86.27% (SD = 17.93) of all components (across the three sessions) and independent coders rated coaches as fully implementing 73.02% of all components, indicating high levels of adherence.

Disagreements between coach and independent coder ratings related most to the determination of a partial versus full implementation and were most notable on (a) items about the explanation of process elements (e.g., whether they explained all steps following the interview; the explanation of the feedback form and linking data elements to the positive behavioral supports and CARES frameworks explicitly; reviewing all steps of the action planning process); (b) the explicit recording of the menu of options during the feedback session; and (c) explicit teacher engagement (e.g., asking for teacher feedback during the feedback session at least once after each of the two feedback form sections). Adherence to aspects related to CARES (i.e., linking feedback data points to the CARES framework and prompting the teacher to set a CARE goal) had notably lower levels of adherence, per both raters, than the adherence for the same items regarding the positive behavioral support elements.

Latent Profiles of Fidelity

We fit a series of models with up to four latent profiles using seven indicators of four components of fidelity: adherence (coach and independent ratings), dosage (total minutes in coaching and number of coach contacts), quality (percent of coach utterances coded as MI consistent), and participant responsiveness (i.e., percent of teacher utterances coded as change talk and teacher rating of the coaching working relationship). Through our review of the fit indices, entropy and posterior probabilities, profile size, and interpretability, we determined that a 2-profile model of fidelity was the best fitting model. The 2-profile model demonstrated a

notable reduction in the BIC, which plateaued for the 3-profile model. The 2-profile model also included notable sample sizes and proportions with clearly interpretable profile differences. Finally, the entropy surpassed 0.80 and approached 0.90 and posterior probabilities were over 0.90. Although the LMRT and VLMR *p*-values were non-significant for the two-profile model, multiple other indices and values indicated support for two profiles. See Table 3 for fit statistics and Figure 1 for a graphical depiction of this model.

Our sensitivity analyses revealed that adjusting for significant correlations (i.e., five significant associations) resulted in a statistically significant improvement in the model ($T = 51.55$, $df = 5$, $p < .01$). This included WITH statements between the coach-rated adherence with external-coder rated adherence, total contacts, and teacher rating of the working relationship; the external coder-rated adherence with teacher rating of the working relationship; and the total minutes spent in coaching with the total number of contacts. Therefore, these WITH statements were retained in all LPA and auxiliary analyses.

We identified the two profiles as having *high* (i.e., above average) *fidelity* and *lower* (below average) *fidelity* (see Figure 1). The high-fidelity profile is characterized by above average fidelity across all measured components. The adherence ratings are high by both raters but are notably discrepant, where coaches reported that they "fully implemented", on average, 90.67% of all components but independent coders reported that 73.52% of components, on average, were fully implemented. The dosage for teachers in this high-fidelity profile was over double as many minutes of coach time (i.e., 441 min versus 174 min on average and 9.82 contacts with the teacher versus 4.18). Participant responsiveness, as measured by teacher change talk was substantially higher (i.e., about 7.55% of teacher utterances were coded as change talk, as compared to 0.11% in the lower-fidelity profile) as were teacher ratings of the

working relationship (i.e., $M = 3.77$ on a scale of 0–4, indicating ratings of *often* to *always*, as compared to a $M = 2.50$, indicating ratings of *sometimes* to *often* among teachers in the lower-fidelity profile). The lower-fidelity profile demonstrated stronger agreement between raters on adherence (i.e., about 63% and 67%) and in fact, the independent coders had higher ratings of adherence than the coaches for the teacher coaching cases in this profile. Interestingly, MI quality was more comparable in the two profiles where MI-consistent language comprised 7.41% of coach utterances among the cases with lower fidelity overall and 10.50% of utterances among the cases with high fidelity overall.

Outcome Analyses

The outcomes of interest included (a) five ASSIST tallies of teacher behaviors (i.e., proactive behavior management, approvals, disapprovals, reactive management, and opportunities to respond [OTRs]), (b) ASSIST global ratings of culturally responsive practices and positive behavioral supports, and (c) three CLASS-S composites measuring emotional support, classroom organization, and instructional support. Student behavior outcomes were measured using the ASSIST tallies of student non-cooperation and disruptive behaviors. The R3STEP analyses revealed only one significant difference in the odds that a teacher was in the high- or lower- fidelity profile based on baseline data; those with better Instructional Support scores on the CLASS were more likely to be in the high-fidelity profile.

The DU3STEP analyses indicated no statistically significant differences between teacher practices for teachers in the high-fidelity and lower-fidelity profiles (see Table 4). However, there were a couple of teacher practices, as measured by the ASSIST, where there were moderately-sized differences (i.e., > 0.30), including OTRs ($d = 0.30$) and reactive behavior

management (d = -0.34). Similarly, there were no significant differences between the two profiles on ASSIST global ratings or the CLASS-S composite scores.

Regarding student behaviors, students within classrooms of teachers in the high-fidelity profile exhibited statistically significantly fewer instances of non-cooperation (M = 0.62) in 15 min than those in classrooms of teachers in the lower fidelity profile (M = 2.12), reflecting a large effect size (d = -0.87; See Table 4).

Discussion

This is a unique study of implementation fidelity, whereby we assessed 4 out of 5 components of fidelity identified by Dane and Schneider (1998), using a range of data sources and reporters, the latter of which has been identified as an area of weakness in the literature, whereby recipients of an intervention are often not included as reporters for engagement/ responsiveness (Lakind et al., 2021). Other unique contributions of this paper were the examination of whether coaches were reliable in self-reporting adherence and highlighting areas of discrepancy for practice implications, as well as examining how the multidimensional measurement of fidelity associated with important teacher practice and student behavior outcomes. Fidelity measurement, in any form, remains quite rare and when included, most commonly addresses adherence or attendance/dosage (Lakind et al., 2021; Stormont et al., 2015). We know of few studies that have assessed multiple fidelity indicators and examined associations with outcomes (for exceptions, see Cross et al., 2015 and Sutherland et al., 2015). Moreover, no studies to our knowledge have assessed such a robust set of fidelity indicators and used a latent approach to examine profiles of fidelity in relation to outcomes.

Taken together, our data suggested that the coach ratings of adherence had close to 80% agreement with independent raters and demonstrated high levels of adherence, on average, as

expected. Notably, the greatest discrepancies were between ratings of partial and full implementation. Although coaches were trained to complete the adherence checklist, a more formal and comprehensive codebook was developed for the independent coders. Such a codebook would likely have improved coach reliability in adherence ratings and should be developed and used for any self-reported measure of adherence. In examining the adherence ratings descriptively, specific areas of training and supervision emerged, including the need to ensure that explanations about the coaching process to teachers are adequate. This may be difficult for coaches to reflect on; as experts in the process, they may not realize that their explanations are briefer or less thorough than an outsider (or the teacher) would report. Coaches may have also shortened these explanations to ensure adequate time for discussions during each session; a feeling of being rushed in problem-solving is a challenge for coaching and consultation (for example, see Newman et al., 2017). Another area of lower fidelity rated by independent coders was the writing of a menu of options during feedback. It is possible that coaches did this without verbalizing it; however, this same component was noted as the lowest fidelity item by coaches during earlier developmental work (Pas et al., 2016).

Finally, the CARES framework did not receive as much attention as the positive behavioral supports elements of the coaching. This finding also emerged in an earlier developmental study (Pas et al., 2016) and has seemingly improved with more focused attention to this area within coach supervision. Anecdotally, in some cases, coaches recounted that when there were too many basic classroom management concerns, focusing on CARES was premature and overwhelming for the teachers. More exploration is needed to determine whether there should be a staggered approach, whereby certain foundational positive behavioral supports are addressed prior to a focus on CARES (i.e., suggesting a possibly more tailored approach to

feedback and goal setting). Another possibility is that setting two goals is too burdensome for some teachers and could be an area of exploration (e.g., through surveying or interviewing teachers).

Regarding the profiles of coaching fidelity, we observed a lower-fidelity and high-fidelity profile where there were notable differences across adherence, dosage, MI quality, and teacher responsiveness between these two profiles. Importantly, over 70% of teachers were engaged in high fidelity coaching. In the lower fidelity profiles, the coaches still maintained a comparable degree of MI, adhered to the protocols for over 60% of components, and had four contacts with teachers (on average). About half of the teachers in this profile never engaged with the coach at all. Future research should examine ways to engage these harder to reach teachers (i.e., who will not engage at all) and examine a more tailored approach (e.g., different scripted conversations to overcome resistance).

Finally, the examination of teachers in the high- versus lower-fidelity profiles was the least conclusive aspect of this study. There was one statistically significant and large effect detected on student non-cooperation, whereby there were fewer instances of tallied non-cooperation in classrooms of teachers who were engaged in high-fidelity coaching. There were also a couple of moderate-sized (though statistically non-significant) differences, all favoring the high-fidelity group, for teacher use of opportunities to respond and reactive behavior management. This may suggest that the observational measure is more sensitive to changes in student behaviors than teacher practices, or that there may be teacher practices that changed and were not measured. Earlier main effects findings for the Double Check model in elementary and middle schools indicated a more robust set of impacts on teacher practices (Bradshaw et al., 2018). The current randomized trial was focused only on middle schools and main effects

analyses are underway. Further research is needed to examine the best ways to measure teacher practice and student behavior changes in response to coaching, including whether there are specific aspects of fidelity that are particularly important for outcomes.

Limitations and Future Research

Although this is a novel study of fidelity of coaching and MI, it does not leverage the randomized controlled trial (RCT) design and therefore causality cannot be inferred. Given sample size limitations, a more complex modeling approach (e.g., BCH in M*plus*) was not conducted, as parameters to estimate exceeded the sample size. The DU3STEP approach is well-supported and robust for our specific research questions that were focused on differences in observed teacher practice and student behavior outcomes (Asparouhov & Muthén, 2013). A larger future study would allow for a more comprehensive analytic approach using BCH and would also be better powered for examining these outcomes than we were in the context of this research, where just over 40 teachers were in the lower-fidelity profile. Although unique in the number of fidelity indicators we measured, Dane and Schneider (1998) identified a fifth component (i.e., exposure/program differentiation) that we did not measure. Additional research should replicate and expand our approach.

The quality of implementation was measured in this study as the use of MI-consistent language but could be considered more broadly to encompass other quality indicators. As has been noted in prior research, fidelity indicators can be challenging to define and capture (Perepletchikova et al., 2007; Schoenwald et al., 2011) and this is perhaps most true for quality (Dane & Schneider, 1998). Only feedback sessions were coded as these sessions had no scripted MI and thus was the best measure of coaches' unscripted use of MI-consistent language (see Pas et al., 2021). In turn, we used these same sessions to code change talk; however, rates of change

talk may be lower in this session relative to others and therefore may not represent true change talk rates for the entire coaching process. Coding of all stages of the coaching process is another area for future research. Finally, although the high fidelity noted was a strength of our efficacy trial, our fidelity levels likely exceeded those of regular coaching practice. Therefore, we only identified a high- and lower-fidelity profile; some research suggests (Newman et al., 2017) that it is likely that additional fidelity profiles would be observed elsewhere and is an area for future research. Furthermore, high levels of fidelity also made a more nuanced examination of specific coaching sessions less feasible.

Implications for School Psychology Practice and Research

This study highlights how coaching fidelity can vary across teachers and how fidelity may play a role in teacher practice and student outcomes, within the context of an RCT with relatively high levels of fidelity. This finding has implications for school psychology research, training and supervision, and practice. Additional research is needed to replicate these findings in a larger sample, with greater fidelity variability, and ideally with other coaching and consultation models to determine if the findings are generalizable. Fidelity measurement is underutilized and when used, is mostly focused on descriptively presenting fidelity levels achieved (Stormont et al., 2015). Fidelity measurement should be integrated further into RCTs and used substantively to address important implementation science research questions about key components of coaching and consultation and the promotion of outcomes (Gottfredson et al., 2015).

Regarding training, supervision, and practice implications, ensuring that a coach is clear and thorough in presenting process information and fully and regularly engages the teacher in collaborative discussion are areas in need of attention; this has been noted in other research summarizing consultee-focused consultation (Newman et al., 2017). Furthermore, the lower

fidelity with addressing CARES may indicate coach hesitance to address cultural responsivity that either stems from or contributes to teacher hesitance. Training and supervision focused on cultural responsivity is essential for school psychologists and is a noted gap in school consultation courses (Hazel et al., 2010). Finally, explicitly considering and collecting data about fidelity is a key data-based decision-making step that needs more attention in practice. The findings here suggest that coaching practitioners could reliably report on adherence but need formal training and clarity on when something is not implemented or is partially or fully implemented. This is something trainers and practitioners could develop or seek out from model developers and utilize.

Conclusions

Coaching may play an important role in promoting changes in teacher practices and student outcomes. This study demonstrated that those teachers receiving high-fidelity coaching also engaged in moderately more OTRs and less reactive behavior management and that students presented fewer instances of non-cooperation. Findings suggest ways to optimize coach training and supervision, ways to ensure coaches may reliably report on adherence, and future research directions.

References

Ackerman, S. J., & Hilsenroth, M. J. (2003). A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review, 23*, 1–33. https://doi.org/10.1016/S0272-7358(02)00146-0.

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—secondary. *School Psychology Review*, *42*(1), 76–98. https://doi.org/10.1080/02796015.2013.12087492.

Amrhein, P. C., Miller, W. R., Yahne, C. E., Palmer, M., & Fulcher, L. (2003). Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of Consulting and Clinical Psychology, 71*(5), 862–878. https://doi.org/10.1037/0022-006X.71.5.862

Apodaca, T. R., Jackson, K. M., Borsari, B., Magill, M., Longabaugh, R., Mastroleo, N. R., & Barnett, N. P. (2016). Which individual therapist behaviors elicit client change talk and sustain talk in Motivational Interviewing? *Journal of Substance Abuse Treatment, 61*, 60–65. https://doi.org/10.1016/j.jsat.2015.09.001

Asparouhov, T., & Muthén, B. (2013). Auxiliary variables in mixture modeling: A 3-step approach using Mplus. *Mplus Web Notes, 15*(6), 1–39. Retrieved from https://www.statmodel.com/examples/webnotes/AuxMixture_submitted_corrected_webnote.pdf

Bishop, D. C., Pankratz, M. M., Hansen, W. B., Albritton, J., Albritton, L., & Strack, J. (2014). Measuring fidelity and adaptation: Reliability of an instrument for school-based

prevention program. *Evaluation & the Health Profession, 37*(2), 231–257. https://doi.org/10.1177/0163278713476882

Borden, L. (2012). Project Arches: An evaluation of a modified Family Check-Up intervention in an assessment setting. Unpublished dissertation. *University of Missouri*.

Bradshaw, C. P., Pas, E. T., Bottiani, J. H., Debnam, K. J., Reinke, W. M., Herman, K. C., & Rosenberg, M. S. (2018). Promoting cultural responsivity and student engagement through double check coaching of classroom teachers: An efficacy study. *School Psychology Review, 47*(2), 118–134. https://doi.org/10.17105/SPR-2017-0119.V47-2

Cohen, J. (Ed.). (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). American Psychological Association.

Cox, J. R., Martinez, R. G., & Southam-Gerow, M. A. (2019). Treatment integrity in psychotherapy research and implications for the delivery of quality mental health services. *Journal of Consulting and Clinical Psychology, 82*(3), 221–233. https://doi.org/10.1037/ccp0000370

Cross, D., Shaw, T., Hadwen, K., Cardoso, P., Slee, P., Roberts, C., Thomas, L., & Barnes, A. (2015). Longitudinal impact of the cyber friendly schools program on adolscents' cyberbullying behavior. *Aggressive Behavior, 42*(2), 166–180. https://doi.org/10.1002/ab.21609

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*, 23–45. https://doi.org/10.1016/S0272-7358(97)00043-3

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*(2), 245–260. https://doi.org/10.1023/A:1019637632584

Durlak, J. A., & DuPre. E. P. (2008). Implementation matters: A review of the research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41,* 327–350. https://doi.org/10.1007/s10464-008-9165-0

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Erchul, W. P., & Sheridan, S. M. (2008). *Handbook of research in school consultation*. Lawrence Erlbaum Associates.

Frey, A., Lee, J., Small, J. W., Sibley, M., Owens, J., Skidmore, B., Johnson, L, Bradshaw, C. P., Moyers, T. (2021). Mechanisms of motivational interviewing: A conceptual framework to guide practice and research. *Prevention Science*. https://doi.org/10.1007/s11121-020-01139-x

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up

research in prevention science: Next Generation. *Prevention Science, 16*, 893–926. https://doi.org/10.1007/s11121-015-0555-x

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge University Press.

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

Hazel, C. E., Laviolette, G. T., & Lineman, J. M. (2010). Training professional psychologists in school-based consultation: What the syllabi suggest. *Training and Education in Professional Psychology, 4*(4), 235–243. https://doi.org/10.1037/a0020072

Heaton, K. J., Hill, C. E., & Edwards, L. A. (1995). Comparing molecular and molar methods of judging therapist techniques. *Psychotherapy Research, 5*, 141–153. https://doi.org/10.1080/10503309512331331266.

Herman, K.C., Reinke, W., & Frey, A. (2020). *Motivational interviewing in schools: Strategies for engaging parents, teachers, and students* (2nd ed.). Springer.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*, 88–110. https://doi.org/10.1080/19345740802539325.

Johnson, S. R., Pas, E. T., & Bradshaw, C. P. (2016). Understanding and measuring coach-teacher alliance: A glimpse inside the 'Black Box'. *Prevention Science, 17*, 439–449. https://doi.org/10.1007/s11121-016-0633-8.

Joyce, B., & Showers, B. (1980). Improving inservice training: The message of research. *Educational Leadership, 37*(5), 379–385.

Lakind, D., Bradley, W. J., Patel, A., Chorpita, B. F., & Becker, K. D. (2021). A multidimensional examination of the measurement of treatment engagement implications for children's mental health services and research. *Journal of Clinical Child & Adolescent Psychology.* https://doi.org/10.1080/15374416.2021.1941057

Larson, M. Cook, C. R., Brewer, S. K., Pullmann, M. D., Hamlin, C., Merle, J. L., Duong, M., Gaias, L., Sullivan, M., Morrell, N., Kulkami, T., Weeks, M., & Lyon, A. R. (2021). Examining the effects of a brief, group-based motivational implementation strategy on mechanisms for teacher behavior change. *Prevention Science*. https://doi.org/10.1007/s11121-020-01191-7

Laws, M. B., Magill, M., Mastroleo, N. R., Gamarel, K. E., Howe, C. J., Walthers, J., Monti, P. M., Souza, T., Wilson, I. B., Rose, G. S., & Kahler, C. W. (2018). A sequential analysis of motivational interviewing technical skills and client responses. *Journal of Substance Abuse Treatment, 92*, 27–34. https://doi.org/10.1016/j.jsat.2018.06.006

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778.

Lyon, A. R., Cook, C. R., Locke, J., Davis, C., Powell, B. J., & Waltz, T. J. (2019). Importance and feasibility of an adapted set of implementation strategies in schools. *Journal of School Psychology, 76*, 66–77. https://doi.org/10.1016/j.jsp.2019.07.014

Magill, M., Gaume, J., Apodaca, T. R., Walthers, J., Mastroleo, N. R., Borsari, B., & Longabaugh, R. (2014). The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. *Journal of Consulting and Clinical Psychology, 82*(6), 973–983. https://doi.org/10.1037/a0036833

Magill, M., Apodaca, T. R., Borsari, B., Gaume, J., Hoadley, A., Gordon, R. E. F., Tonigan, J. S., & Moyers, T. (2018). A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology, 86*(2), 140–157. https://doi.org/10.1037/ccp0000250

Martin, T., Moyers, T. B., Houck, J., Christopher, P & Miller, W. R. (N. D.). *Motivational Interviewing Sequential Code for Observing Process Exchanges (MI-SCOPE) Coding Manual.* University of New Mexico, Center on Alcoholism, Substance Abuse and Addictions (CASAA). Unpublished manual. https://casaa.unm.edu/download/scope.pdf

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. Little (Ed.), *The Oxford handbook of quantitative methods* (pp. 551–611). Oxford University Press.

Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change* (3rd ed.). Guilford Press.

Moyers, T. B., & Martin, T. (2006). Therapist influence on client language during motivational interviewing sessions. *Journal of Substance Abuse Treatment*, *30*(3), 245–251. https://doi.org/10.1016/j.jsat.2005.12.003.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 345–368). Sage.

Muthén, L. K., & Muthén, B. O. (1997–2012). *Mplus user's guide*. Muthén and Muthén.

Mutschler, C., Naccarato, E., Rouse, J., Davey, C., & McShane, K. (2018). Realist-informed review of motivational interviewing for adolescent health behaviors. *Systematic Reviews, 7,* 109–129. https://doi.org/10.1186/s13643-018-0767-9

Nagin, D. S. (2005). *Group-based modeling of development over the life course*. Harvard University Press.

Newman, D. S., McKenney, E. L. W., Silva, A. E., Clare, M., Salmon, D., & Jackson, S. (2017). A qualitative metasynthesis of consultation process research: What we know and where to go. *Journal of Educational and Psychological Consultation, 27*(1), 13–51. https://doi.org/10.1080/10474412.2015.1127164

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Owens, J. S., Lee, M., Kassab, H., Evans, S. W., & Coles, E. C. (2021). Motivational Ruler Ratings among Teachers Receiving Coaching in Classroom Management: Measurement and Relationship to Implementation Integrity. *Prevention Science.* https://doi.org/10.1007/s11121-020-01111-9

Pas, E. T., Borden, L., Herman, K., & Bradshaw, C. P. (2021). Leveraging Motivational Interviewing to coach teachers in the implementation of preventive evidence-based practices: A sequential analysis of the Motivational Interviewing process. *Prevention Science.* https://doi.org/10.1007/s11121-021-01238-3

Pas, E., & Bradshaw, C.P. (2021). Introduction to the special issue on optimizing the implementation and effectiveness of preventive interventions through motivational interviewing. *Prevention Science, 22*(6), 683–688.

Pas, E. T., Duran, C. A. K., Debnam, K. D., & Bradshaw, C. P. (2022). Is it more effective or efficient to coach teachers using traditional one-on-one coaching or in teacher pairs?: A comparison of impacts on teacher and student outcomes. *Journal of School Psychology.*

Pas, E. T., Larson, K., Reinke, W., Herman, K., & Bradshaw, C. P. (2016). Implementation and acceptability of an adapted Classroom Check-Up coaching model to promote culturally responsive classroom management. *The Education & Treatment of Children, 39*(4), 467–491. https://doi.org/10.1353/etc.2016.0021

Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: Survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology, 77*(2), 212–218. https://psycnet.apa.org/doi/10.1037/a0015232

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*, *75*(6), 829–841. https://doi.org/10.1037/0022-006X.75.6.829

Pianta, R. C., Hamre, B., Hayes, N., Mintz, S., & LaParo, K. M. (2008). *Classroom Assessment Scoring System–Secondary* (CLASS-S). University of Virginia.

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing science*, *12*(1), 103–124. https://doi.org/10.1287/mksc.12.1.103

Reinke, W. M., Lewis-Palmer, T., & Merrell, K. (2008). The Classroom Check-Up: A classwide teacher consultation model for increasing praise and decreasing disruptive behavior. *School Psychology Review, 37*(3), 315–322.

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/10.1007/s10488-010-0321-0

Rusby, J. C., Crowley, R., Sprague, J., & Biglan, A. (2011). Observations of the middle school environment: The context for student behavior beyond the classroom. *Psychology in the Schools*, *48*(4), 400–415. https://doi.org/10.1002/pits.20562.

Rusby, J. C., Taylor, T. K., & Milchak, C. (2001). Assessing school settings: Interactions of students and teachers. *Eugene, OR: Oregon Research Institute*.

Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*, 243–248.

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

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Sheridan, S. M., Swanger-Gagne, M., Welch, G. W., Kwon, K., & Garbacz, S. A. (2009). Fidelity measurement in consultation: Psychometric issues and preliminary examination. *School Psychology Review, 38*(4), 476–495.

Snape, L., & Atkinson, C. (2016). The evidence for student-focused motivational interviewing in educational settings: A review of the literature. *Advances in School Mental Health, 9* (2), 119–139. http://dx.doi.org/10.1080/1754730X.2016.1157027

Solomon, B. G., Klein, S. A., & Politylo, B. C. (2012). The effect of performance feedback on teachers' treatment integrity: A meta-analysis of the single-case literature. *School Psychology Review*, *41*(2), 160–175. https://doi.org/10.1080/02796015.2012.12087518

Stahmer, A. C., Rieth, S., Lee, E., Reisinger, E. M., Mandell, D. S., & Connell, J. (2015). Training teachers to use evidence-based practices for autism: Examining procedural implementation fidelity. *Psychology in the Schools, 52* (2), 181–195. https://doi.org/10.1002/pits.21815

Stormont, M., Reinke, W. M., Newcomer, L., Marchese, D., & Lewis C. (2015). Coaching teachers' use of social behavior interventions to improve children's outcomes: A review of the literature. *Journal of Positive Behavior Interventions, 17*(2), 69–82. https://doi.org/10.1177/1098300714550657

Sutherland, K. S., Conroy, M. A., Vo, A., & Ladwig, C. (2015). Implementation integrity of practice-based coaching: Preliminary results from the BEST in CLASS efficacy trial. *School Mental Health, 7,* 21–33. https://doi.org/10.1007/s12310-014-9134-8

Sutherland, K. S., Conroy, M. A., McLeod, B. D., Algina, J., & Wu, E. (2018). Teacher competence of delivery of BEST in CLASS as a mediator of treatment effects. *School Mental Health, 10,* 214–225. https://doi.org/10.1007/s12310-017-9224-5

Tapp, J., & Walden, T. (1993). PROCODER: A professional tape control, coding, and analysis system for behavioral research using videotape. *Behavior Research Methods, Instruments, & Computers*, *25*(1), 53–56. https://doi.org/10.3758/BF03204449

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. E. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of*

Consulting and Clinical Psychology, 61, 620–630. https://doi.org/10.1037/0022-006X.61.4.620

Wasserman, L. (1997). *Bayesian model selection and model averaging*. Presented at Mathematical Psychology Symposium "Methods for Model Selection," Bloomington, IN.

Table 1

Sample demographics

	Full intervention sample (n = 153)		Completed coaching (n =128)	
Teacher characteristics	Total	%	Total	%
Female gender	138	90.2%	119	93.0%
Race/Ethnicity				
White	70	45.8%	56	43.8%
Black/African American	56	36.6%	52	40.6%
Asian/Pacific Islander	7	4.6%	5	3.9%
Hispanic/Latino	1	0.7%	1	0.8%
Other	9	5.9%	8	6.3%
Age				
20-30	39	25.5%	32	25.0%
31-40	39	25.5%	34	26.6%
41-50	29	19.0%	27	21.1%
51-60	18	11.8%	13	10.2%
61+	8	5.2%	6	4.7%
Years Teaching at This School				
1st Year	29	19.0%	27	21.1%
1-3 Years	41	26.8%	36	28.1%
4-8 Years	47	30.7%	36	28.1%
9+ Years	22	14.4%	19	14.8%
Grade Taught				
Sixth	52	34.0%	44	34.4%
Seventh	43	28.1%	33	25.8%
Eighth	12	7.8%	9	7.0%
Multiple	46	30.1%	42	32.8%
Subject Taught				
English/Language Arts	41	26.8%	34	26.6%
Social Studies/History	38	24.8%	32	25.0%
Math	49	32.0%	36	28.1%
Science	43	28.1%	38	29.7%
Design Variables				
Cohort 1	34	22.2%	27	21.1%
Cohort 2	38	24.8%	35	27.3%
Cohort 3	36	23.5%	30	23.4%
Cohort 4	45	29.4%	36	28.1%

Note. Because of missing data, not all totals add to the sample size and precents do not always equal 100%. No significant differences between samples on demographics.

Table 2

Baseline ASSIST and CLASS-S scores

Measure	*M*	*SD*	Range
ASSIST teacher practice tallies			
Proactive behavioral management	9.57	4.29	2.00-27.33
Approval	3.28	3.37	0.00-22.00
Disapproval	0.80	0.99	0.00-5.00
Reactive behavior management	7.67	4.65	1.33-29.67
Opportunities to Respond (OTR)	15.72	9.44	0.33-60.00
ASSIST teacher practice global ratings			
Culturally responsive practices	1.18	0.51	0.39-2.76
Positive behavioral supports	0.91	0.28	0.43-2.00
CLASS-S composites			
Emotional support	4.25	0.79	2.44-6.67
Classroom organization	3.86	0.54	2.22-5.00
Instructional support	3.46	0.79	1.27-5.53
ASSIST student behavior tallies			
Non-cooperation	1.50	2.32	0.00-12.33
Disruptive behavior	15.02	13.99	0.00-79.00

Note. ASSIST = Assessing School Settings: Interactions of Students and Teachers and CLASS = Classroom Assessment Scoring System –Secondary version. The ASSIST tallies had no cap in occurrences; the *culturally responsive practices* scale items were scored on a scale of 0–4, the *positive behavioral supports* scale items were scored on a scale of 0–2, and the CLASS-S dimensions were scored on a scale of 1–7. All data are the averages of three observations and therefore the ranges of tallies do not always reflect whole numbers.

Table 3

Latent Profile Analyses results

Classes	Log likelihood	AIC	BIC	A-BIC	Entropy	LMR	VLMR	Class sizes	Post. Prob.
1	-1101.41	2040.823	2098.151	2038.018				151 (100%)	
2	**-864.56**	**1797.12**	**1899.71**	**1792.11**	**0.917**	**0.12**	**0.12**	**44 (29.1%)** **107 (70.9%)**	**0.96** **0.98**
3	-831.317	1746.634	1873.359	1740.434	0.923	0.24	0.24	43 (28.5%) 92 (60.9%) 16 (10.6%)	0.97 0.97 0.97
4	-808.552	1717.103	1867.967	1,709,723	0.898	0.37	0.36	45 (29.8%) 11 (7.3%) 81 (53.6%) 14 (9.3%)	0.95 0.99 0.94 0.95

Note. AIC = *Akaike Information Criteria*; BIC = *Bayesian Information Criterion*; A-BIC = *sample size adjusted BIC*; LMR = *Lo-Mendell-Rubin likelihood ratio test*; VLMR = *Vuong- likelihood ratio test*; and Post. Prob. = *Posterior Probabilities*. All indices were standardized. WITH statements were included to allow for covariance between statistically significantly correlated variables within the LPA. BF was < 1 and cmP was 0 for all models.

Table 4

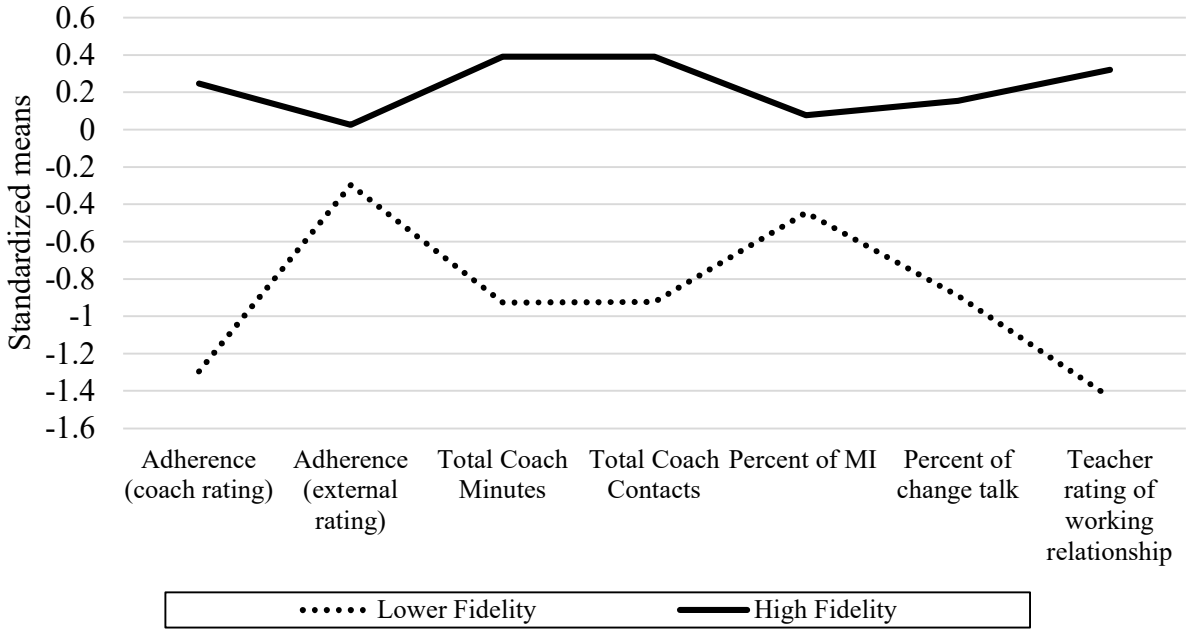
Auxiliary analyses examining mean differences between profiles on outcomes

		Lower Fidelity		High Fidelity				
		M	(*SE*)	*M*	(*SE*)	χ^2	*p*-value	*d*
Teacher Practices (ASSIST Tallies)	Proactive Behavior Management	7.97	0.43	8.89	0.51	1.79	0.18	0.26
	Approvals	2.54	0.41	3.10	0.33	1.31	0.25	0.24
	Disapprovals	0.66	1.27	0.80	0.30	0.01	0.93	0.13
	Reactive Behavior Management	7.68	1.11	6.24	0.33	1.48	0.22	**-0.34**
	OTRs	14.07	1.26	16.69	1.15	2.33	0.13	**0.30**
Teacher Practices (ASSIST Globals)	Culturally Responsive Teaching	1.15	0.11	1.13	0.04	0.01	0.92	-0.03
	Positive Behavior Supports	0.77	0.05	0.82	0.04	1.72	0.19	0.23
Teacher Practices (CLASS composites)	Emotional Support	4.85	0.30	4.66	0.13	0.42	0.52	-0.22
	Classroom Organization	4.02	0.13	3.87	0.10	1.71	0.19	-0.25
	Instructional Support	3.89	0.32	3.69	0.12	0.38	0.54	-0.22
Student Behaviors (ASSIST Tallies)	Non-cooperation	2.12	0.65	0.62	0.16	4.43	**0.04**	**-0.87**
	Disruptive Behavior	9.71	2.69	11.49	1.35	0.39	0.53	0.18

Note. ASSIST = *Assessing School Settings*; Interactions of Students and Teachers and CLASS = *Classroom Assessment Scoring System – Secondary version*. χ^2 = chi-square test; *d* = Cohen's *d* effect size. There were no significant differences on any of these outcomes between classes at baseline. Bold indicates significant *p*-value and a moderate or large effect size.

Figure 1

Latent Profile Analysis Plot for Fidelity



Note. The sample sizes were 44 (29.1%) for the lower-fidelity profile and 107 (70.9%) for the high-fidelity profile.