Summer 7-1-2021

A Longitudinal Analysis of Trajectories and Predictors of Fidelity Using the SafeCare Parenting Model

Matthew J. Lyons

A Longitudinal Analysis of Trajectories and Predictors of Fidelity Using the SafeCare Parenting Model

By

Matthew Lyons

A Dissertation

Submitted to the School of Public Health

Georgia State University

In partial fulfillment of the requirements

For the degree of Doctor of Philosophy in Public Health

July 2021

Contents

Chapter 1. Statement of Purpose and Literature Review

Statement of Purpose

Evidence-based interventions are increasingly used in public health, mental health, and child welfare settings in an effort to improve services and outcomes for families. Even if an evidence-based intervention has been demonstrated in clinical trials to have a positive impact on consumer outcomes, it must be implemented appropriately to be fully effective (Aarons et al., 2011; Bauer & Kirchner, 2019). The specificity and relatively strict requirements of evidence-based practice (EBP), however, are often at odds with the challenging context and emergent nature of clinical work in various practice settings. There is a growing body of literature studying the factors that impact implementation success or failure, as well as identifying various measurable domains of implementation. One such domain is fidelity, which has been defined as, "the degree to which an intervention was implemented as it was prescribed in the original protocol or as it was intended by the program developers" (Proctor et al., 2011). Implementation fidelity has been shown to be associated with consumer outcomes in numerous studies, and though results across the implementation literature are somewhat mixed (Chiapa et al., 2015; Forgatch & DeGarmo, 2011; Schoenwald et al., 2000; Smith et al., 2013), there is broad agreement that fidelity is a critical implementation outcome.

In practice, efforts to ensure high fidelity, such as coaching and monitoring, have been concentrated in early implementation phases, with support generally waning as time goes along. While many studies have examined initial indicators of implementation success, far fewer have examined what happens over time. That is, there are limited studies that have tracked implementation fidelity trajectories over time, and even fewer have analyzed the predictors of fidelity trajectories. Information is limited regarding whether implementation fidelity tends to increase, decrease, or stabilize over time, and what individual characteristics might be associated with those outcomes. **The purpose of the present study** is to use data from the ongoing implementation and dissemination of the SafeCare model to better understand 1) the characteristics of SafeCare implementation fidelity trajectories, and 2) factors that predict differences in those trajectories. This analysis will contribute to the empirical literature on the sustainability of EBP implementation, and help inform how fidelity monitoring and coaching practices should change over time.

Review of the Literature

This literature review will begin with a brief overview of the SafeCare model, which is the focus of the current investigation, followed by a discussion of the field of implementation science and those domains of implementation that have been identified and measured (with a particular focus on fidelity). Next, I will present a summary of the findings from several studies that have analyzed implementation fidelity longitudinally, drawing from the education literature, from rehabilitation/physical therapy, and from child welfare. I will then briefly discuss a key individual level factor that may relate to implementation fidelity, provider attitudes toward EBP, as well as the measure through which that construct is operationalized in this study: the Evidence Based Practice Attitude Scale (EBPAS) (Aarons, 2004). Finally, the present study will be introduced in more detail, and the ways in which it represents a novel contribution to the literature on EBP implementation fidelity trajectories will be outlined.

SafeCare

SafeCare is a behavioral parenting program used for child maltreatment prevention and early intervention targeting parents with children ages 0 to 5. The SafeCare curriculum is made up of 3 modules: Health, Home Safety, and Parent Child/Parent Infant Interaction. The Health module provides parents with structured steps to identify injuries and illnesses in children and to determine the appropriate course of action, whether that be treating the child at home, consulting a physician, or contacting emergency services. The Home Safety module provides parents with information on how to keep a hazard free home environment and teaches them about the importance of appropriate supervision. The Parent Child Interaction module provides parents with skills to manage their interactions with their children in healthy and nurturing ways, to structure their activities together, and to manage problem behaviors in children 18 months to 5 years old. The Parent Infant Interaction module provides similar guidance, but for children less than 18 months old.

The SafeCare curriculum is made up of eighteen sessions, with six sessions for each module. Within each module, the six sessions have a similar structure. The first session focuses on a pre-assessment of the skills addressed in the module, while the second through fifth sessions focus on training parents in those specific skills. The sixth session focuses on a post-training skill assessment to gauge skill mastery. Throughout the training process, SafeCare providers impart skills based on four major principles: explain, model, practice, and feedback. In other words, the SafeCare providers explain the skills that are the focus of that day's session, demonstrate those skills for the parents, give the parents opportunities to practice those skills, and then provide constructive feedback for the parents based on their performance of those skills.

As part of standard SafeCare implementation, fidelity is assessed by SafeCare coaches or trainers who either accompany providers on their sessions, or more commonly, listen to audio recordings of the sessions once completed. These sessions are then scored for fidelity using standardized rating scales that include items that assess important process variables (e.g., building rapport, structuring the session) and content variables (e.g., explaining the skills, modeling a skill), and providers are given feedback on their session with the goal of increasing fidelity over time. This process is termed ‘coaching’ in the NSTRC’s implementation model. The frequency of coaching is front-loaded so that newly-trained providers are coached more frequently (e.g., weekly) than experienced providers who have demonstrated high fidelity to each of SafeCare’s modules and achieved certification. The specifics of SafeCare’s fidelity measures are discussed in greater detail in Chapter 2: Methods, and the measures themselves are provided in Appendix 1.

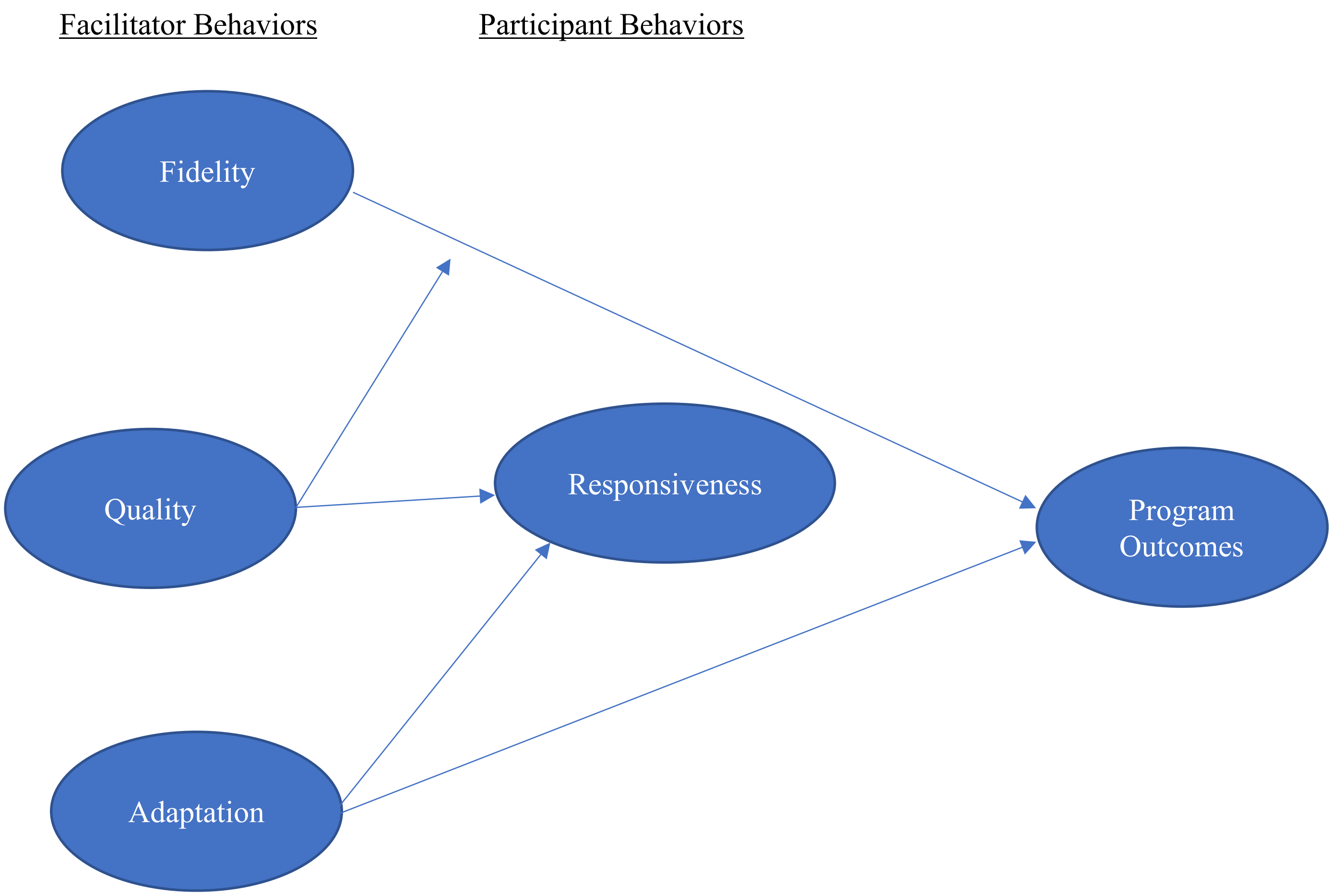
Implementation Science

The effectiveness of EBP’s in real world practice is dependent upon successful implementation (Fixsen, 2005). Over the last several decades, efforts to better understand the mechanisms that contribute to success or failure in the implementation of EBP’s has led to the development of implementation science as an independent area of scientific inquiry. Implementation science has been defined as “the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice, and, hence, to improve the quality and effectiveness of health services and care” (Eccles & Mittman, 2006, p. 1). While other areas of research in health and human services appropriately focus on end-user outcomes, implementation research is distinguished by its focus on outcomes directly related to program implementation. Within the implementation science literature, there are various theoretical frameworks that conceptualize implementation components and outcomes, the relationships between these components and outcomes, and the relationships between implementation outcomes and end-user outcomes (Aarons et al., 2011; Damschroder et al., 2009; Fixsen, 2005).

Proctor et al. (2011) identify a range of implementation outcomes, including acceptability, adoption, appropriateness, costs, feasibility, penetration, sustainability, and fidelity. Each of these terms comes with a degree of interpretive complexity, and there is some diversity in the definitions applied across the literature. However, a few basic definitions (adapted from Proctor et al, 2011) that are particularly relevant for this study are as follows: *acceptability* refers to stakeholder opinions on the given treatment; *adoption* can be understood as the degree of uptake of an EBP within a practice setting; *appropriateness* is the "perceived fit, relevance, or compatibility of the innovation or evidence-based practice for a given practice setting," (Proctor et al, 2011, p. 69); and *fidelity* is the degree to which the implementer follows the protocol outlined by the EBP developers. Together, these and other implementation outcomes are understood to influence service-related outcomes such as efficacy, effectiveness, and equity. These service-related outcomes are thought to directly impact end-user outcomes, such as health behavior, parenting skill, or educational attainment that are the ultimate target of EBP implementation.

While Proctor et al. (2011) focus mainly on definitions of key implementation terms, Berkel et al. (2011) put forth a model that conceptualizes the *relationships* between various implementation components. They separate the components into two categories. These are: "behaviors of program facilitators (fidelity, quality of delivery, and adaptation) and behaviors of participants (responsiveness)" (Berkel et al., p. 23). Within their conceptual model (a version of which is presented in Figure 1 below), implementation components interact to produce program outcomes, with fidelity, adaptation (i.e., situationally or culturally appropriate modifications), and client responsiveness each exerting a direct effect. It is important to note that, while this theoretical framework provides a useful basis for understanding implementation components and outcomes, there are numerous possible ways to conceptualize implementation components and their relationships, each of which may have certain advantages and disadvantages.

Figure 1. Theoretical model of implementation components and outcomes[1]



While there are several plausible theoretical frameworks for understanding the factors that contribute to program outcomes and end-user outcomes, this much is clear: successful implementation is a necessary precondition for program effectiveness (Durlak & DuPre, 2008). Further, successful implementation is dependent on a variety of factors both intrinsic and extrinsic to implementing organizations. The EPIS model, put forth by Aarons et al. (2011), is

---

[1] Adapted from Berkel et al. (2011).

another important conceptual model for understanding the factors that influence implementation. Within this framework, implementation is understood as unfolding in four phases: Exploration, Preparation, Implementation, and Sustainment. Within each of these phases, there are “inner context” and “outer context” factors that can impact the success or failure of an implementation. One set of inner context factors that influence implementation are the characteristics of the individual providing the intervention. Particularly during the Implementation and Sustainment phases, individual staff attitudes toward evidence-based practice are identified as an important factor. As will be discussed in greater detail below, the relationship between these staff attitudes and implementation fidelity constitutes a central theoretical interest and analytic focus of the study undertaken here.

Implementation fidelity: definition and measurement

Defining implementation fidelity. Dusenbury et al. (2003) define fidelity as the extent to which service providers “implement programs *as intended by the program developers*” (p. 240). While this definition is similar to the one shared above from Proctor et al. (2011) and seems quite straightforward, there is still a degree of inconsistency in the implementation science literature regarding the operationalization and measurement of this core construct. For example, while the Berkel et al. (2011) model above conceptualizes fidelity as separate from service quality, Proctor et al (2011) interpret quality as a component of implementation fidelity itself. To add to this complexity, there is some diversity in the definition of service quality in the literature as well. For example, Berkel et al. (2011) include in their discussion of quality of delivery some components that are arguably covered within the SafeCare fidelity rating scale, such as interactive teaching methods, but also include other components that are not covered therein, such as enthusiasm of delivery. Despite this complexity, the consensus in the implementation science literature is that fidelity is an important implementation component and that implementers should strive for high fidelity to ensure program effectiveness.

Measuring implementation fidelity. Given the critical importance of appropriate measurement in ensuring service quality and advancing implementation research, there have been significant efforts to systematize the theoretical models and measurement instruments used to conceptualize implementation components in recent years (Lewis et al., 2015). However, there are some obstacles to these efforts in the case of fidelity. Fidelity is generally measured either as a set of behaviors that were completed or not (as is the case with SafeCare), or as the "amount of time dedicated to each of the core components," of an EBP (Berkel et al., 2011, p. 25). Due to the fact different EBP's have different core components, there are no generalizable measures of fidelity that can be used across all programs. While some EBP's have a broad enough base of dissemination and accompanying fidelity monitoring for researchers to conduct systematic investigation into the strengths and weaknesses of the various fidelity measurement instruments (i.e. Muse & McManus, 2013), most EBP's are disseminated on a smaller scale and by a smaller number of implementers. For EBP's like SafeCare that are disseminated by a single institution through a network of partner agencies, there is only one fidelity instrument in use across these partner agencies at any given time. However, this instrument contains items that are specific to the SafeCare program, and therefore there can be no reasonable expectation of direct comparison of the instrument with those used for other EBP's. Further, fidelity data collection procedures vary widely. This diversity is underscored in Bartley et al. (2017) review of fidelity predictors in child welfare settings, in which: "Fidelity data were gathered through self-report in five studies … client report in four studies … observation in four studies … a combination of self-report and client report in two studies … therapist report, client report, and observation in one study … and case notes in the remaining study" (p. 438). Among these different approaches, the scholarly consensus is generally that expert ratings of fidelity should be the preferred approach (Schoenwald et al., 2004).

Implementation fidelity: predictors and outcomes

What predicts implementation fidelity? Several studies have explored the relationships between individual and organizational factors and fidelity. One recent structured literature review covering 15 research studies assessed how individual and organizational characteristics impact the fidelity of interventions in child welfare settings (Bartley et al. 2017). It found some inconsistencies in the literature regarding individual level predictors of fidelity, though at the organizational level implementation of continuous quality improvement and coaching practices were consistently related to high fidelity across several studies (Schoenwald et al., 2004; Bearman et al., 2013; Webster-Stratton et al. 2014). Even though the literature is less consistent regarding predictors of fidelity at the individual level, there are still a range of such factors that may impact implementation fidelity. The review by Bartley et al. (2017) found that age, sex, and practice experience were each related to fidelity in at least one study, though in their words, "the results were mixed" (p. 438). Bartley et al. (2017) found one study in which younger age was associated with higher fidelity, but after adjustments for multiple testing those results were no longer significant (Whitaker et al., 2011). Bearman et al., (2013) found that sex was not statistically significantly related to fidelity, but Beidas et al., (2015) found that, for one of three interventions under study, female sex was predictive of higher fidelity. Bartley et al. (2017) found similarly mixed results in the literature with regard to work experience and fidelity: Beidas et al. (2015) found a significant negative association with work experience and fidelity, with more time in the field corresponding to a lower likelihood of implementation fidelity, while

Taylor et al., (2015) found almost precisely the opposite. Importantly, provider attitudes toward evidence-based practice may be related to fidelity, as providers who find a particular evidence-based intervention appealing are often more willing to use that intervention (Reding et al., 2014). The relationship between provider attitudes and fidelity is often complex, however, as positive provider attitudes toward evidence-based practice have been shown to be associated with both fidelity-consistent and fidelity-inconsistent program modifications in practice (Wiltsey Stirman et al., 2015). In other words, providers with positive attitudes toward EBP may be inclined to make changes in implementation that are not consistent with the prescribed curriculum.

Implementation fidelity and outcomes. Over the course of the last several decades, there has been a growing body of literature exploring the relationship between implementation fidelity and program/end-user outcomes. Botvin et al. (1989) found that implementation fidelity was associated with program efficacy in a skills-based smoking prevention curriculum among Latinx youth. In a study of the impact of early interventions among children in the first grade to prevent later substance abuse, Ialongo et al. (1999) found that higher levels of program fidelity were associated with higher behavior ratings and achievement scores (the key outcomes of the study) among participating children. Durlak and DuPre (2008) conducted a review of over 500 studies on implementation components and outcomes, and found that, in the subset of studies where implementation components were explicitly assessed in relation to outcomes, level of implementation was positively associated with program outcomes in in 45 out of 59 studies. In most cases, fidelity was a key implementation component measured within these studies, and the authors state that, “minimal variability in implementation levels could be an explanation for the weak or null results obtained in 8 of the remaining 14 studies” (Durlak & DuPre, 2008, p. 331). Implementation fidelity has also been shown in multiple studies to be related to program and end-user outcomes specifically among children and families (e.g.Forgatch & DeGarmo, 2011; Chiapa et al., 2015). Since these studies also track fidelity longitudinally, they will be discussed in greater detail in the next section on fidelity trajectories. In addition to the evidence that high fidelity is associated with positive program and end-user outcomes, there is some evidence to suggest that fidelity *monitoring* activities themselves can exert a positive influence on outcomes over and above fidelity, and this influence has been demonstrated in the context of SafeCare implementation. Aarons et al. (2009) showed that, when coupled with fidelity monitoring, the implementation of SafeCare actually resulted in a protective effect *against* staff turnover. In other words, those service providers who implemented SafeCare and received fidelity monitoring and coaching support, were less likely to leave their agencies than those implementing services as usual (with or without coaching support) and those implementing SafeCare without support.

Implementation fidelity trajectories and the impacts of fidelity over time

While implementation fidelity is an increasingly studied topic, most studies focus on initial implementation outcomes: do providers adopt a model at all and deliver it with fidelity? As important, however, is what happens over time. Do providers continue to maintain fidelity to a model, or do they drift when purveyor support begins to diminish? Further, how do fidelity trajectories impact program and end-user outcomes?

Fidelity trajectories. There are few studies that have examined fidelity longitudinally, and even fewer have focused on predictors of implementation fidelity trajectories. Among those that do track fidelity longitudinally, the majority of which are from the education literature, the findings are somewhat mixed regarding the nature of fidelity trajectories. Schaper et al. (2016) studied the within-year growth in 353 schools' fidelity to the School-Wide Positive Behavioral Interventions and Supports (SWPBIS) intervention using the Team Implementation Checklist

(TIC). Predictors included school-level variables such as year of implementation (1-4), size of student body, degree of urbanity, and SES (measured as proportion of students eligible for free or reduced price lunch). No individual or lower level variables were measured. Multilevel growth models found that rural schools, smaller schools, elementary/middle schools, and schools in years 3 and 4 of implementation had statistically significantly higher scores at the beginning of the year than their comparators. Fidelity tended to increase over time, and growth (change in fidelity per month of school) was predicted by year of implementation and SES. Growth was highest in year two, with a hypothesized ceiling effect accounting for slower growth in years 3 and 4. Lower SES was associated with lower fidelity growth rates.

Another study from the education literature used multilevel growth models to study fidelity trajectories in the implementation of two early childhood literacy programs in 100 classrooms across 52 elementary schools (Zvoch, 2009). Fidelity was measured three times over two weeks using observational checklists completed by direct classroom observations, and overall trajectories varied between sites. Results indicated that both within and between school variation predicted changes in adherence, and fidelity trajectories differed between the two implemented programs, with an overall increase in fidelity for one and an overall decrease for the other. Hoekstra et al. (2017) reported findings from a study analyzing longitudinal fidelity data from a Dutch physical activity promotion program in multidisciplinary rehabilitation care settings. They used hierarchical cluster analysis to analyze data on ~70 professionals across 17 sites at 3 separate time points. These analyses identified three separate fidelity trajectory profiles: "stable high fidelity (n = 9), moderate and improving fidelity (n = 6), and unstable fidelity (n = 2)" (Hoekstra et al., 2017). Smaller implementation sites and sites that adopted the intervention early were more likely to be in the stable high fidelity category. Further, sites where providers reported high degrees of support from clinical staff, professional appreciation, and positive assessments of program fit were more likely to have stable high fidelity.

McIntosh et al. (2016) identified similar trajectory profiles using latent class analysis of School Wide Positive Behavioral Interventions and Supports implementation in 5331 schools over a 5 year period. Four latent classes were identified, with two representing successful implementation (sustainers and slow starters), and two representing failed implementation (late abandoners and rapid abandoners). Successful implementation was predicted by grade level served (favoring elementary schools), school size (favoring larger schools), and having a greater number of schools in a given district participating in the implementation process. Another study on PBIS showed in a randomized trial including 58 high schools that those schools with higher baseline rates of bullying implemented PBIS with higher fidelity over the course of two-years, indicating that the context and specific challenges of the implementing organization may increase motivation to implement with fidelity over time (Bradshaw et al., 2015). One additional study from the education literature followed 23 teachers over a 3 year period to measure their fidelity to a drug prevention curriculum (Ringwalt et al., 2010). In this instance, fidelity was measured by videotaping sessions and having trained coders score each observation for fidelity. Their "brute force" multilevel models showed a significant degree of heterogeneity within teachers with respect to fidelity, and no identifiable pattern was observed save for a regression to the mean.

Forgatch and Degarmo (2011) present the results of three studies on the national implementation of the Parent Management Training – Oregon Model™ (PMTO) in Norway, two of which involve longitudinal analysis of fidelity. The first study followed 35 implementers of PMTO, 29 of whom completed the study and were included in analyses. Video recordings of

PMTO sessions were assessed at three time points (early, mid, and late) and assessed for fidelity using the Fidelity of Implementation Rating Scale (Knutson et al., 2009). The researchers hypothesized that fidelity would tend to increase from one time point to the next, and that variability would decrease. Study one findings supported these hypotheses, as fidelity improved across all domains of the measurement instrument (knowledge, structure, teaching, process, and overall development). The second study explored whether fidelity was sustained when training activities were transferred from PMTO's purveyors to the implementing community organizations. The researchers followed three generations of providers (G1, G2, and G3), where G1 providers were trained by the purveyors and the subsequent generations were trained by community practitioners. Analyses found a small decline in fidelity from G1 to G2 but no observable decline when comparing G3 to G1, and the authors determined that their hypothesis (that fidelity would tend to decrease when training activities were transferred into the community) was not supported.

The study which most closely parallels that which is conducted here was published by Chaffin et al. (2016). The authors present the findings of a SafeCare implementation study comparing fidelity trajectories between 9 providers on intensively trained Interdisciplinary Collaborative Teams (ICT's) and 36 providers from subsequent cohorts who were trained by the ICT members. Overall, providers in the study served a total of 957 clients in 5,769 individual sessions. The authors defined fidelity as, "adherence to basic behaviors prescribed by the model, and not necessarily the expertise with which a behavior is executed" (Chaffin et al., 2016). Fidelity was measured through a client report questionnaire with two sections: one focused on the general style of service delivery and another checklist of specific provider behaviors.

To track fidelity over time across different SafeCare modules, the following fidelity domains were identified, measured, and factor analyzed as indicators of the latent fidelity construct: psychoeducation, teaching/modeling, feedback, homework, and resources. Their main predictor of interest was a cohort variable at the provider level representing membership in either the ICT seed team or a subsequent cohort. Using a multi-level growth modeling approach, the authors found that fidelity tended to improve over time. There were statistically significant intercepts between ICT members and non-members, indicating that those who were members of ICT's had higher initial post-training fidelity. However, they did not find any differences in fidelity slopes, and within one year, fidelity levels between ICT and non-ICT cohorts were practically identical. While gender and race were tracked and analyzed at the provider level, no significant effects on fidelity were observed.

The impacts of fidelity over time. While the focus of the present study is on fidelity trajectories and their predictors, and not the relationships between those trajectories and other outcomes, it is important to interpret this discussion in the context of those latter relationships. Much like the literature on fidelity trajectories themselves, studies relating fidelity over time to other outcomes yield mixed results. In a recent study on PBIS implementation that analyzed both educational and behavioral outcomes among students in 85 Ohio schools over the course of 2 years, improvements in implementation fidelity (as measured by a tiered fidelity inventory) were associated with reduced rates of out-of-school suspensions, but not with changes in academic achievement (James et al. 2019). Another, larger study of the same intervention, however found that among the 477 participating schools across ten US states, "there were no significant associations between fidelity and change in behavior or academic outcomes" (Kim et al., 2018). The results of this study were complicated, however, by the fact that 78% of participating schools met minimum fidelity criteria, so the comparisons made were generally between "high fidelity" schools, and those schools which were merely "at fidelity," rather than between those with high vs. low fidelity. Further, among schools with sustained implementation, office discipline referrals and out-of-school suspensions decreased over a three-year period. The authors suggest that these findings indicate that outcomes related to extremely high fidelity may not differ significantly to outcomes related to merely sufficient fidelity.

Chiapa et al. (2015) present the results of a randomized trial including families receiving services through the Women, Infants, and Children Nutritional Supplement Program (WIC). Seventy-nine families with children aged two years who exhibited problem behaviors participated in the study, which sought to investigate: 1) four-year fidelity trajectories among therapists implementing a structured parenting skills intervention targeting youth adjustment (The Family Check Up – FCU), and 2) the relationship between fidelity trajectories and youth outcomes, specifically: "oppositional and aggressive behaviors at ages 7.5 and 8.5 years" (Chiapa et al., 2015, p. 1007). Using latent growth curve modeling, the authors found that fidelity tended to decrease linearly over time, and that "steeper declines were related to less improvement in caregiver-reported problem behaviors" (Chiapa et al., 2015, p. 1006). The authors take these findings to indicate that continuous fidelity monitoring and support is a critical component in implementation to ensure sustained fidelity and positive end-user outcomes. However, the interpretation of these findings are complicated by Clements et al. (2015), who, in the context of a randomized trial, studied the implementation of an early mathematics curriculum among 64 teachers in 26 schools in low-income districts and found that teacher fidelity to the curriculum was stable, high, and continuing to increase even 2 years after the cessation of fidelity supports. From these conflicting results, we can at least conclude that further study is warranted, both on the relationships between fidelity trajectories and end-user outcomes, and on the characteristics and predictors of those trajectories themselves.

Attitudes toward evidence-based practice

As noted above, individual provider attitudes toward evidence-based practice have an important, if complex, relationship with implementation fidelity (Reding et al., 2014; Wiltsey Stirman et al., 2015). In an effort to measure and understand these provider attitudes, Aarons (2004) developed the fifteen item Evidence Based Practice Attitude Scale (EBPAS-15) which is a measurement tool used to identify provider attitudes toward evidence-based practice with four subdomains: requirements, appeal, openness and divergence. The requirements subdomain assesses how willing providers would be to adopt an evidence-based practice if it was required by their employer, while the appeal subdomain assesses provider perceptions of the intuitive appeal of evidence-based practice. Openness assesses provider willingness to follow an evidence-based practice curriculum, and divergence assesses the degree to which providers are resistant to employing evidence-based interventions. The EBPAS-15 (which is utilized in this study), has also been expanded into a 50-item version with 12 sub-domains (Aarons et al., 2012), and subsequently truncated again to include only 36 items, but the same number of sub-domains (Rye et al., 2017)

In a 2006 study on the relationship between organizational culture and EBP attitudes (measured by the EBPAS-15), Aarons and Sawitzky (2006) found in correlational analyses and hierarchical regression analyses that “constructive” organizational cultures were associated with more positive EBP attitudes, while “defensive” organizational cultures were associated with more negative EBP attitudes. In this case, constructive cultures “are characterized by organizational norms of achievement and motivation, individualism and self-actualization, and being humanistic and supportive," while defensive cultures "are characterized by seeking approval and consensus, being conventional and conforming, and being dependent and subservient" (Aarons & Sawitsky, 2006, p. 62). Aarons et al. (2010) conducted an investigation into the psychometric properties of the EBPAS scale in a sample of 1,089 US-based mental health service providers. In this study, they also analyzed the relationships between demographic characteristics and EBPAS scores, and found that female sex, Caucasian race, and education in social work were related to higher EBPAS scores.

There is some indication that EBP attitudes as measured by EBPAS scales are related to implementation fidelity, though again the results are mixed. As noted above, while Reding et al (2014) found that EBP attitudes were related to provider willingness to implement EBP's, Wiltsey Stirman et al. (2015), found that even positive attitudes are associated with both fidelity-consistent and fidelity- inconsistent modifications to EBP curricula. Both of these studies utilized the EBPAS as the key attitudinal measure. Further, in a recent study of a behavioral intervention for children with autism spectrum disorder (pivotal response treatment), therapists' general attitudes toward EBP were not significantly associated with fidelity, however, openness to innovation (measured on a different scale), *was* related to fidelity of implementation, "indicating that therapists who were more willing to use or try new interventions and EBPs had higher levels of fidelity" (Verschuur et al., 2019, p. 506). In the context of a randomized implementation trial of cognitive processing therapy for post-traumatic stress disorder including 78 clinicians, two EBPAS subscales (openness and requirements) were significantly correlated with a measure of implementation fidelity (the Therapist Adherence and Competence Rating Form), with a correlation of .25 between openness and adherence and .27 between requirements and adherence (Sijercic et al., 2020). In this same study, regression analyses indicated that the openness and requirements factors accounted for 11% of the variability in fidelity (termed adherence), though while "requirements significantly predicted TAC adherence … openness did not" (Sijercic et al., 2020, p. 13). Thus, though there is some suggestion in the literature that fidelity may be positively associated with provider attitudes toward EBP, the results are somewhat inconsistent.

Limitations of the existing literature

While a limited number of other studies have tracked EBP implementation fidelity longitudinally, and one study has done so in the context of SafeCare implementation, there are significant gaps in the empirical understanding of fidelity trajectories and their predictors. Certain unique characteristics of the data set in the present study have made it possible to begin to fill some of those gaps. While some studies have shown implementation fidelity to generally improve over time, others have identified several classes of fidelity trajectory, ranging from the successful and consistently improving to outright implementation failure (which can be understood to indicate zero percent fidelity). Among those studies that have identified these distinct classes, none provided insight as to the individual level predictors that might be associated with different fidelity outcomes or trajectories. For example, while McIntosh et al. (2016) analyzed data from a SWPBIS implementation in over 5,000 schools, fidelity was measured at the school level, and no individual level variables were measured as potential predictors of implementation fidelity. The current dataset contains a range of individual level variables that are absent in the larger studies described above. Further, while some large-scale studies have been conducted that provide insight into fidelity longitudinally, in child welfare those studies have been limited in scope. Chaffin et al. (2016) followed SafeCare providers very closely, but their implementation was limited to a small geographic region and a small group of providers at a limited number of agencies. In addition, Chaffin et al. (2016) measured fidelity via client report rather than via expert judgment, which is the preferred method (Mowbray et al., 2003). To my knowledge, there is no study in a child welfare context that analyzes individual level predictors of EBP implementation fidelity over time in a sample this large, with data that extends over a number of years.

The contribution of this study

Through ongoing implementations, the National SafeCare Training and Research Center (NSTRC), which disseminates the SafeCare model, has collected 14,778 fidelity observations of SafeCare sessions by some 868 providers in 172 agencies. This data set, collected over 10 years, contains longitudinal observations on implementation fidelity among SafeCare providers, as well as a range of individual provider variables that may be related to fidelity, including demographics, work history, and a key attitudinal measure, the Evidence-Based Practice Attitude Scale (Aarons, 2004). The current project uses data collected from ongoing SafeCare implementations to address several research questions. In addition to a broad exploration of whether fidelity tends to increase, decrease, or stabilize over time, I analyze the relationships between provider level variables (demographics and attitudes toward evidence-based practice) and change in fidelity over time, with a particular focus on how attitudes toward evidence-based practice impact fidelity over time. Thus, the unique contributions of the proposed study may be summarized as follows: 1) an empirical evaluation of fidelity trajectories among EBP providers that is larger and more geographically diverse than any in the child welfare implementation literature, and 2) an analysis of the individual level variables that may be associated with differential implementation fidelity trajectories, with a particular focus on the effect of provider attitudes toward EBP.

Chapter 2: Methods

Data Source and Collection Procedures

The National SafeCare Training and Research Center disseminates SafeCare nationally and internationally. Organizations and governments contract with NSTRC to provide SafeCare training and implementation support to providers who will implement SafeCare. When a new SafeCare provider is registered for SafeCare training, they are asked to complete a brief registration form which includes the provider level predictor variables to be analyzed in this study. These data are gathered and stored in the SafeCare training portal. Once provider trainees complete workshop training, providers receive in-field “coaching” in which certified SafeCare coaches either accompany the provider or listen to audio recordings of sessions. Coaches then score individual sessions for fidelity using SafeCare fidelity checklists. Coaches can be housed either at NSTRC or at the implementing agency. For new agencies, NSTRC typically conducts coaching until an agency is ready to assume that function. All providers who complete registration forms are asked whether they consent to have the associated data used for research purposes, and only those data associated with providers who answer yes are included in this research. In total, 868 (90.2%) providers consented and 95 (9.8%) of providers did not consent.

Data Structure

The data set generated through the process described above has a hierarchical structure, and may be understood as having four distinct levels of nesting (or clustering). Individual session observations are nested within providers, who are nested both within coaches (raters/scorers) and within organizations (sites). Coaches are not necessarily nested within sites, however, as providers are often coached initially by trainers housed at NSTRC before being transitioned to coaches housed within their home agency. To render parameter estimates interpretable, it is preferable to select a maximum of three levels of nesting to represent during multilevel model estimation. While there are informative variables available at the provider level, there is no unique descriptive information available at either the coach or organization level. Theoretical interest was greater in the variance that may be explained by nesting at the organizational level, and the imperfect hierarchical structure in which individual providers may be nested within multiple coaches in both NSTRC and partner agencies would present practical and analytic challenges. Organization was therefore selected as the third level to include in our modeling process. A visual representation of the nesting structure of the data as analyzed is provided in Figure 2 below.

Figure 2. Representation of the nesting structure of the data as analyzed

Instruments and Measures/Variables of Interest

Level 1: Session level variables

SafeCare Fidelity: SafeCare fidelity is the outcome of interest, and is measured for each session by a trained observer (a SafeCare coach or trainer) who completes a rating scale with items that are the required elements for the session. There are three types of sessions across the three SafeCare modules: pre-assessment/baseline, training, and post-assessment/end of module. For each type of session, a slightly different scale is used. In each case, between 27 and 32 items are rated by the observer as either having occurred or not (with a third possibility of "not applicable"). Fidelity score, the operationalization of the construct of interest (SafeCare fidelity), is constituted in the number of behaviors that occurred divided by the total number of behaviors that should have occurred (i.e., occurred/(occurred + did not occur); items rated not applicable are excluded from the computation). As such, the fidelity measure used with the SafeCare model is conceptually restricted in comparison with the expansive definition provided by Proctor et al. (2011), which includes factors such as dosage and quality of delivery as components of fidelity. Checklists for each SafeCare module can be found in Appendix 1.

Observation time: For each fidelity observation, the number of days since the first observation is coded, with the first observation coded as zero days. This critical time variable allows us to model fidelity trajectories, as well as predictor influences on fidelity trajectories, using time by predictor interactions.

Level 2: Provider level predictors of fidelity

Provider race: Provider race is available in the data set, and is used in the inferential modeling process. Originally, this variable had six levels: African American, American Indian, Asian American, White, Hispanic, and Other. Due to small sample size within several of those categories, the version of the variable tested in inferential analyses was reduced to a binary white/non-white variable.

Provider sex: Provider sex is available in the data set and is tested in the modeling process.

Years on the job: The number of years on the job that a provider is available as a continuous variable, and is included in the modeling process.

Provider academic degree: Provider education is measured using a three-level variable indicating the highest degree attained: high school, bachelor's degree, or graduate degree, which is tested in the modeling process.

The EBPAS Scale: The Evidence Based Practice Attitude Scale is a key predictor of interest. While these original subdomains are important, EPBAS scores in this sample have been factor analyzed to determine whether the original proposed factor structure holds, or whether a different structure is warranted. Factor analyses yielded two factors, which I term EBPAS Positivity and EBPAS Negativity. Greater detail on this factor analytic process is provided in Chapter 3.

Experience working with at risk families: The amount of experience a provider has working with at risk families is operationalized as having three levels: none, less than one year, or greater than one year.

Experience working with substantiated families: The amount of experience a provider has working with families with substantiated maltreatment is operationalized as having three levels: none, less than one year, or greater than one year.

Training providers to work with high risk families: The amount of experience a provider has training other providers to work with high risk families is operationalized as having three levels: none, less than one year, or greater than one year.

Experience providing structured parenting interventions: The amount of experience a provider has implementing structured parenting interventions is operationalized as having three levels: none, less than one year, or greater than one year.

Training providers in structured parenting interventions: The amount of experience a provider has training other providers to implement structured parenting interventions is operationalized as having three levels: none, less than one year, or greater than one year.

Prior training in evidence based interventions: Whether a provider has received prior training in evidence-based interventions is operationalized as a binary (yes/no) variable.

Prior training in structured parenting interventions: Whether a provider has received prior training in structured parenting interventions is operationalized as a binary (yes/no) variable.

Preliminary Hypotheses

Hypothesis 1: Overall Trajectories

Fidelity will tend to increase over time until a ceiling is reached, in keeping with the findings of Chaffin et al. (2016).

Hypothesis 2: predictors of Fidelity Trajectories

Baseline fidelity (the fidelity intercept) and fidelity growth (fidelity slopes) may be impacted by individual level variables other than EBP attitudes. These analyses include demographic and work history variables in an exploratory capacity, however, so no specific hypotheses have been made about the nature of those relationships. While a literature search yielded mixed results, I hypothesize, in keeping with the findings of Reding et al. (2014) and Sijercic et al. (2019), that positive provider attitudes toward evidence-based practice, as measured by EBPAS scores, will be associated with higher baseline fidelity. However, in keeping with the findings of Chaffin et al. (2016), I hypothesize that a ceiling effect will preclude the possibility of a positive association between positive EBP attitudes and fidelity *growth*. Instead, because positive attitudes are hypothesized to be associated with higher baseline scores, I hypothesize that providers with positive attitudes will have less room to grow. With regard to negative EBP attitudes, I hypothesize that they will be associated both with lower baseline fidelity scores and lower rates of fidelity growth.

General analytic process

The first analytic step is an in-depth descriptive analysis of univariate and bivariate distributions of the variables of interest. These exploratory analyses provide critical insight into the distributions of fidelity conditional on predictors of interest, and a descriptive sense of whether fidelity tends to increase, decrease, or stabilize over time. I then conduct factor analyses on the evidence-based practice attitude scale to determine appropriate scoring of each measure and any subscales, results of which are reported in Chapter 3.

Upon completion of the exploratory phase, inferential statistical models are constructed. Given the nested and longitudinal structure of the data in question, the model building process will involve constructing a taxonomy of multilevel models (AKA general linear mixed models, random effects models) with observations nested within providers, and providers nested within organizations. This approach will account for the correlation between observations within providers, and allow for the estimation of random slopes and intercepts for provider level predictors of fidelity. Crucially, this multilevel approach will allow for the estimation of time by predictor interaction terms which will constitute the main outcomes for this study. The general analytic approach to testing this project's central hypotheses is described below.

Hypothesis 1: Overall Trajectories

An overall assessment of the tendency for fidelity to increase, decrease, or remain constant over time has been made using a general linear mixed model with fidelity as the dependent variable, time as a fixed effect, and provider as the grouping variable. The beta coefficient associated with time and a null hypothesis test for statistical significance act as indicators of whether significant positive or negative fidelity trajectories can be detected. In an effort to identify the appropriate functional form of the relationship between time and fidelity, non-linear functions of time will also be explored.

Hypothesis 2: Predictors of Fidelity Trajectories

2.1, Attitudes: EBPAS scores are included in a general linear mixed model as level 2 (provider level) predictors. Cross-level interaction terms for EBPAS variables and time are calculated to determine the effect of evidence-based practice attitudes on fidelity over time.

2.2 Work History: All work history variables are included in a general linear mixed model as level 2 (provider level) predictors. Cross-level interaction terms for work history variables are calculated to determine the effect of work history on fidelity over time.

2.3 Demographics: Demographics are included in a general linear mixed model as level 2 (provider level) predictors. Cross-level interaction terms for demographic variables are calculated to determine the effect of demographics on fidelity over time.

Subjects and Setting

The analytic data set contains 14,778 session observations nested within 868 providers in 172 sites. Just over 55% percent of the sample was white, while 93.43% was female. 85.48% of providers were full time, and 57.60% had a bachelor's degree but no graduate degree. The mean number of sessions per provider was 17 (SD = 21.30), while the median number was 13. Table 1 below provides greater details on the sample makeup.

Table 1. Provider demographics

Categorical Variables (N = 868, missing = 0)	N (%)			
Gender				
Female	811 (93.43)			
Male	57 (6.57)			
Provider Race				
African American	90 (10.37)			
American Indian	26 (3.00)			
Asian American	8 (0.92)			
Caucasian	478 (55.07)			
Hispanic	215 (24.77)			
Other	215 (5.88)			
Provider white (0,1)				
White	478 (55.07)			
Non-white	390 (44.93)			
Provider Education				
High School	157 (18.09)			
Bachelor's Degree	500 (57.60)			
Graduate Degree	211 (24.31)			
Provider Full Time				
Part Time	126 (14.52)			
Full Time	742 (85.48)			
Continuous Variables	**Mean**	**SD**	**Median**	**Min, Max**
Years on the job (N = 791, missing = 77)	3.08	4.76	1.00	0, 31

Chapter 3: Data analysis and Results

In this chapter I will outline in more detail the analytic approach to the central research questions introduced above. I will include a detailed account of the descriptive/exploratory analyses and their respective results, as well as an account of the inferential analyses and their results. I will present model statements in equation form for models that represent crucial steps in the iterative model building process and provide accompanying interpretations of the model parameters represented therein. The results of these models, as well as numerous intermediate models, will be reported in table form and interpreted in the text. My focus in this chapter will be on the technical description and statistical interpretation of the model results, and I will defer most discussion of practical significance to Chapter 4: Discussion.

Descriptive analyses

The descriptive (or exploratory) phase is a crucial first step in any statistical analysis. My analytic process begins with the assessment of univariate distributions (some of which are reported in the sample description in Chapter 2 above) of variables of interest (individual level predictors, as well as the session level fidelity outcome), and the consultation of appropriate visual representations of those distributions. I also report the process and findings of the factor analysis on the EBPAS measure in this section. After assessing univariate distributions and the EBPAS factor structure, I assess select joint distributions of variables of interest.

Univariate distributions of variables of interest

Univariate distributions of provider-level fidelity variables. Summary statistics for fidelity variables are reported in Table 2 below. Fidelity score represents the percentage of checks on a SafeCare fidelity checklist multiplied by 100. For each session type (Assessment, Training, and End of Module) there are slightly different checklists. Across 14,778 sessions conducted by

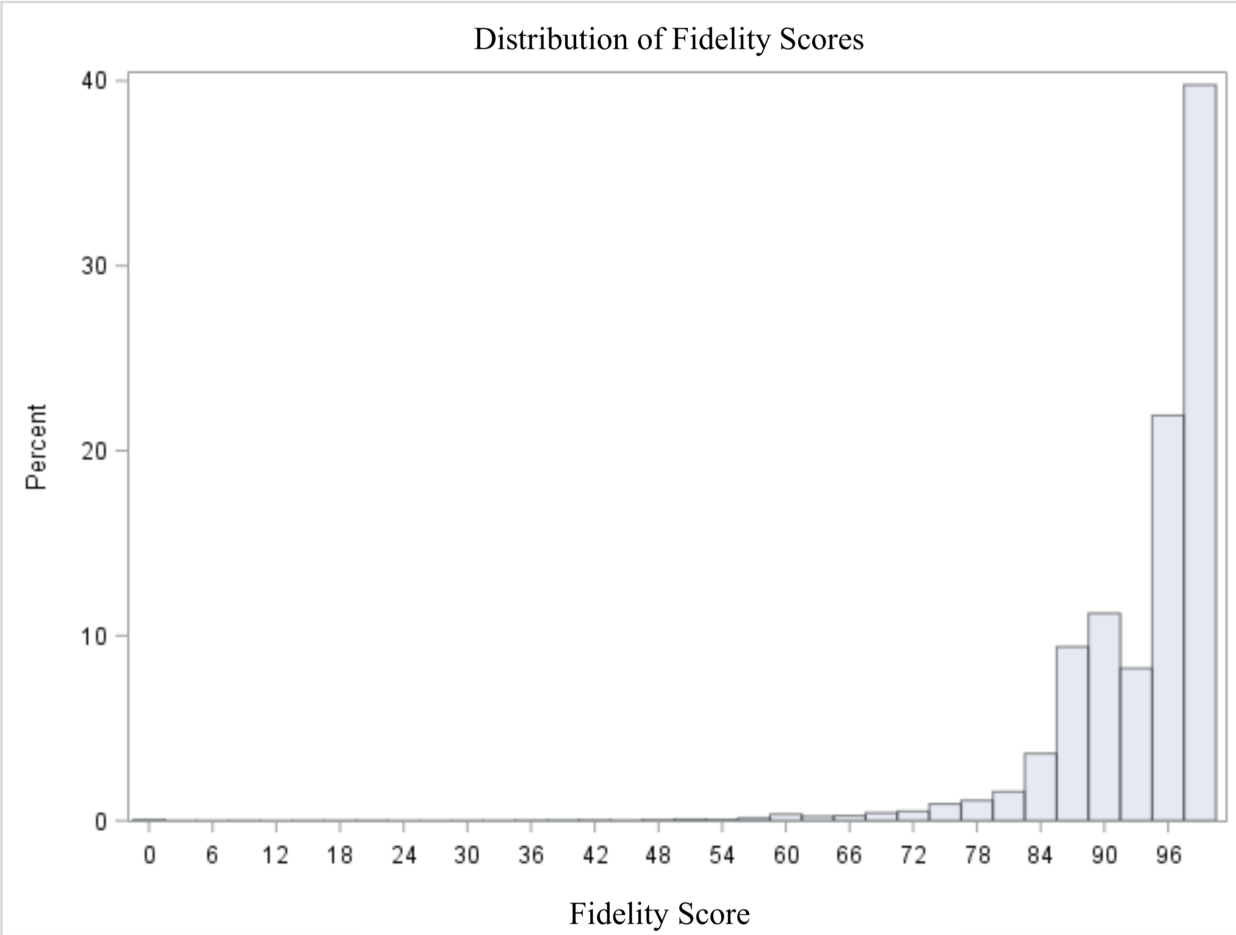
868 providers, the mean fidelity score was 93.93 (SD = 7.90), with a median score of 95.65. Among all sessions observed, 13,612 (92.11%) reached the pass threshold of 85%.

Table 2. Univariate descriptive statistics for fidelity score.

Continuous Variables	N (14,778) (1 missing)	Mean	SD	Median	Min, Max
All sessions	14778	93.93	7.90	95.65	0, 100
Assessment Sessions	4297	94.48	7.33	95.65	0, 100
Training Sessions	9437	93.72	8.03	95.83	0, 100
End of Module	1044	93.56	8.79	95.65	0, 100

A histogram representing the distribution of fidelity scores across all sessions is provided in Figure 4 below. The distribution is strongly left skewed, with more than 75% of observations being above 90% fidelity.

Figure 4. Histogram of fidelity scores across all sessions (N = 14,778)



Univariate distributions of provider-level service experience variables. Provider service experience variables were originally collected with 6 levels (None, 0-6 months, 6-12 months, 1-3 years, 3-5 years, >5 years). To aid in parameter interpretation in inferential models, and due to the small number of observations within certain levels, these variables were recoded as 3 level variables (None, <1 year, >1 year). Out of 868 providers, 662 (76.27%) reported having more than one year of experience working with at-risk families, while a slightly smaller proportion (68.66%) reported having more than one year of experience with substantiated families. The majority of providers had no experience training others to work with high risk families (63.82%) or in structured parenting interventions (66.94%). Most providers had at least some prior experience providing structured parenting interventions, with 44.01% having more than 1 year of experience and 23.62% having less than one year. In addition to provider service experience, providers were also asked two questions regarding their training exposure. Most providers reported never having received prior training in evidence based interventions (63.94%) or in structured parenting programs (54.49%). More detailed information about provider service experience and training can be found in table 3 below.

Table 3. Provider service experience variables (3 level variables)

Variables ((N = 868, missing =0)	**N (%)**
Experience w/ at risk families	
(0) None	69 (7.95)
(1) <1 year	137 (15.78)
(2) >1year	662 (76.27)
Experience w/ substantiated families	
(0) None	102 (11.75)
(1) <1 year	170 (19.59)
(2) >1year	596 (68.66)
Training providers to work w/ high risk families	
(0) None	554 (63.82)
(1) <1 year	89 (10.25)
(2) >1year	225 (25.92)
Exp providing structured parenting interventions	
(0) None	281 (32.37)
(1) <1 year	205 (23.62)
(2) >1year	382 (44.01)
Training Providers in structured par. Interventions	
(0) None	581 (66.94)
(1) <1 year	108 (12.44)
(2) >1year	179 (20.62)
Prior training in structured parenting program	
Yes	395 (45.51)
No	473 (54.49)
Prior training in EBI	
Yes	313 (36.06)
No	555 (63.94)

Univariate distributions of provider-level attitudes toward evidence-based practice. The primary predictor of interest, provider attitudes toward evidence-based practice, was assessed using the Evidence-Based Practice Attitude Scale (EBPAS) described in Chapter 1. Among the 15 items in this scale, the distributions suggest overall positive attitudes toward evidence based practice, with item 1, for example, having a median score of 4 out of 5. In this case, 4 indicates agreement "to a great extent," with the statement, "I like to use new types of therapy/interventions to help my clients." For items that indicate positive EBP attitudes, measures of central tendency tend toward agreement, and for items that indicate negative EBP attitudes, these measures tend toward disagreement. More detailed information on provider attitudes toward evidence-based practice is presented in Table 4 below.

Table 4. Univariate distributions of individual EBPAS items

Variable	Item	Mean	SD	Median
EBPAS1 (N = 851, missing = 17) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	I like to use new types of therapy/interventions to help my clients	3.99	0.81	4.00
EBPAS2 (N = 854, missing = 14) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	I am willing to try new types of therapy/interventions even if I have to follow a treatment manual	4.22	0.79	4.00
EBPAS3 (N = 849, missing = 19) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	I know better than academic researchers how to care for my clients	1.82	0.94	2.00

EBPAS4 (N = 849, missing = 19) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	I am willing to use new and different types of therapy/interventions developed by researchers	4.15	0.77	4.00
EBPAS5 (N = 848, missing = 20) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	Research based treatments/interventions are not clinically useful	1.54	1.02	1.00
EBPAS6 (N = 838, missing = 30) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	Clinical experience is more important than using manualized therapy/treatment	2.17	0.96	2.00
EBPAS7 (N = 843, missing = 25) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	I would not use manualized therapy /interventions	1.51	0.94	1.00
EBPAS8 (N = 851, missing = 17) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	I would try a new therapy/intervention even if it were very different from what I am used to doing	4.00	0.85	4.00
EBPAS9 (N = 842, missing = 26) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	How likely to adopt if it was intuitively appealing?	4.13	0.80	4.00
EBPAS10 (N = 844, missing = 24) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	How likely to adopt if it “made sense” to you?	4.32	0.75	4.00

EBPAS11 (N = 843, missing = 25) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	How likely to adopt if it was required by your supervisor?	4.27	0.83	4.00
EBPAS12 (N = 844, missing = 24) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	How likely to adopt if it was required by your agency?	4.33	0.80	4.00
EBPAS13 (N = 841, missing =) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	How likely to adopt if it was required by your state?	4.27	0.92	4.00
EBPAS14 (N = 843, missing = 25) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	How likely to adopt if it was being used by colleagues who were happy with it?	4.06	0.86	4.00
EBPAS15 (N = 842, missing = 26) (1) Not at all (2) To a slight extent (3) To a moderate extent (4) To a great extent (5) To a very great extent	How likely to adopt if it was being used by colleagues who were happy with it?	4.40	0.78	5.00

For the EBPAS instrument, a factor analytic process was used to produce summary variables to be included in the final analysis. While the initial factor structure found by Aarons (2004) and reproduced as a first-order factor structure in Aarons et al. (2010) included each of the sub-domains described above (Appeal, Requirements, Openness, and Divergence), that factor structure did not reproduce in this sample. An initial principal components analysis with varimax rotation yielded a solution with three eigenvalues greater than one (5.65, 1.87, and 1.63). The three factor solution, however, was difficult to interpret as multiple items had factor loadings greater than .40 for more than one factor and thus did not load clearly onto a single factor. I examined both one and two factor solutions, and ultimately selected the 2-factor solution which yielded two conceptually coherent factors (the single factor solution simply excluded the items on Factor 2 of the 2-factor solution). The final rotated factor loadings are shown below. Factor 1 is termed "EBPAS Positivity" while factor two is termed "EBPAS Negativity." To create the variables for EBPAS Positivity and EBPAS Negativity the items loading onto each factor were averaged to create the factor score. The mean value for the EBPAS Positivity was 4.21 and the mean value for EBPAS Negativity was 1.74. Factor loadings for each of the two factors are reported in Table 5 below.

Table 5. Rotated factor loadings for the EBPAS instrument.

Item	Positivity	Negativity
ebpas1	0.61	-0.15
ebpas2	0.66	-0.15
ebpas3	-0.05	0.49
ebpas4	0.70	-0.19
ebpas5	-0.06	0.76
ebpas6	0.01	0.59
ebpas7	-0.08	0.72
ebpas8	0.70	-0.20
ebpas9	0.73	-0.14
ebpas10	0.72	-0.11
ebpas11	0.78	0.15
ebpas12	0.78	0.16
ebpas13	0.72	0.18
ebpas14	0.72	0.01
ebpas15	0.71	-0.08

Select joint distributions of variables of interest

In order to explore the relationships between categorical predictors and fidelity outcomes, select joint distributions were assessed. Overall, demographic predictors (sex, race, and education) do not appear to be particularly strongly related to fidelity scores, and the same is true for full time status. It is important to note that in the table below the descriptive statistics for fidelity are calculated across all sessions conducted by providers within a given category. For example, out of 14,778 scored sessions, 13,740 of those sessions were conducted by women. Among these 13,740 sessions conducted by women, the mean fidelity score was 93.94.

Table 6. Fidelity score by provider variables (N = total number of sessions)

Variables	N	Mean	SD	Median	Min, Max
Sex					
Female	13740	93.94	7.90	95.65	0, 100
Male	1038	93.77	7.83	95.65	37.04, 100
Race					
Af. Amer.	1161	93.69	7.34	95.45	20, 100
Amer. Ind.	376	92.98	9.53	95.83	18.52, 100
Asian Amer.	250	97.69	5.06	100.00	62.96, 100
Caucasian	7603	93.99	7.97	95.65	0, 100
Hispanic	4536	93.96	7.71	95.65	0, 100
Other	852	92.84	8.47	95.24	8.33, 100
Race (Binary)					
White	7603	93.99	7.97	95.65	0, 100
Non-White	7175	93.86	7.82	95.65	0, 100
Education					
High School	3209	93.34	8.12	95.45	0, 100
Bachelor’s	8145	93.81	8.06	95.65	0, 100
Graduate	3424	94.78	7.20	96.15	31.82, 100
Full Time					
Full Time	12311	93.92	7.80	95.65	0, 100
Part Time	2467	93.97	8.36	95.65	0, 100

Inferential analyses

The methodological literature on longitudinal data analysis is diverse and complex, with numerous approaches used for modeling similarly structured data (including methods related to structural equation modeling, such as latent growth curve modeling, and extensions of the standard linear regression modeling framework, such as multilevel modeling). There is also a broad array of statistical software packages equipped to conduct these analyses (including R, STATA, MLwiN, Mplus, and SAS). In this case, a multilevel modeling (MLM) approach was used, and all analyses were conducted using SAS Version 9.4. Even within the methodological literature specifically regarding MLM, there is significant diversity. For example, naming conventions for these methods are an issue of some complexity, with terms such as multilevel modeling, hierarchical linear regression, linear mixed modeling, and random coefficients modeling being used to describe analyses that use the same general quantitative procedures to estimate parameters. Complexity arises in more substantive areas as well, however, as different methodological reference texts often suggest somewhat different steps in the model building process, different statistical measures of model fit, different parameter estimation techniques, and different ways of prioritizing theory and data in model selection.

Highlighting this heterogeneity, Snijders and Bosker (2012) state that multilevel model building is a complex process, given that there are two “steering wheels” that modelers must use in building a taxonomy of models and selecting an appropriate final model: “substantive (subject-matter related) and statistical considerations” (p. 102). Which of those two steering wheels should be prioritized is dependent on numerous factors. In iterating a model taxonomy to fit the data under study here, there are many such considerations, including the theoretical importance of specific variables, parsimony, interpretability, and statistical measures of model fit. In the text that follows, I will briefly describe the methodological literature from which I draw my analytic approach, outline and justify my general approach to the modeling process, report and interpret models that represent key steps in the model taxonomy, and interpret the results of the final model yielded through that iterative process.

Key literature

I consulted several key texts in developing and executing my analytic approach. My analyses drew heavily on Snijders and Bosker's (2012) textbook, *Multilevel Analysis*. Their in-depth treatment of the quantitative underpinnings of multilevel models provided the conceptual foundation for the analyses performed here, and their work was particularly useful in determining the appropriateness of multilevel modeling for the research questions and data structure entailed in this project, as well as theoretical and practical considerations in balancing data and theory in model selection. *Multilevel Analysis* was also crucial in my decision-making process regarding the functional form of the relationship between time and fidelity. I follow the notation provided in their text for expressing multilevel models in equation form.

Singer's (2002) "Fitting individual growth models using in SAS PROC MIXED" proved a crucial resource for the statistical programming portion of this project, with a significant amount of the SAS code used for fitting the models reported below directly adapted from the code provided therein. That book chapter provides an in-depth discussion not only of model parameters, but specifically of the output generated by various applications of SAS PROC MIXED to longitudinal data. My substantive interpretation of the model parameters, in particular the interpretation of variance components, was informed directly by Singer's discussion of these matters. My SAS code, reading of SAS output, and parameter interpretations were additionally supplemented by Bell et al's (2013) "A multilevel model primer using SAS PROC MIXED," and

Bell et al's (2014) "An intermediate primer to estimating linear multilevel models using SAS PROC MIXED." Greater interpretive depth for the three-level MLM context was drawn from Suzuki and Sheu's (1999) "Using PROC MIXED in hierarchical linear models: examples from two- and three-level school-effect analysis, and meta-analysis research," and Tasca et al's (2009) "Three-level multilevel growth models for nested change data: A guide for group treatment researchers." Numerous specific recommendations regarding the steps in building a growth model taxonomy in an MLM context were provided in Bliese and Ployhart's (2002) "Growth Modeling Using Random Coefficient Models: Model Building, Testing, and Illustrations," and specific recommendations for the use of cross level interaction effects and interpretation of the associated parameters were provided in Aguinis et al.'s (2013) "Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling." My decisions regarding parameter estimation were further guided by Boedeker's (2017) "Hierarchical linear modeling with maximum likelihood, restricted maximum likelihood, and fully Bayesian estimation."

While each of the texts named here offered crucial guidance in developing and executing my analytic approach, as mentioned above there is some degree of diversity in the specific recommendations given across this body of literature. No single text that I could discover provided comprehensive instructions for approaching the research questions and data structure in this project (i.e. some texts addressed multilevel models for longitudinal data, but only for two levels of nesting, while others addressed three-level data structures but not in a longitudinal context). Even Snijders and Bosker's fairly comprehensive textbook devotes only one half of one chapter to longitudinal data structures with a variable occasions design, as is the case in these data, as opposed to designs in which observations occur at fixed intervals for all subjects. Where possible, I have erred in cases of disagreement in the literature on the side of Singer's seminal 2002 work, as the applications outlined in that article map most closely onto my own research.

Why MLM?

Why MLM rather than standard linear regression or ANOVA with Repeated Measures? MLM was selected over standard linear regression for several reasons. First, the data analyzed have a hierarchical (clustered, nested) structure, with session observations nested within providers nested within sites (ignoring clustering within coach). Failure to account for this three-level nesting structure could lead to biased standard error estimates and an increased type 1 error rate (Snijders & Bosker, 2012; Tasca et al, 2008). Relatedly, given that the research questions of interest are longitudinal in nature (as session observations clustered within providers and organizations occur over time), a model which regards session observations as nested within providers is both conceptually and quantitatively appropriate. Growth models (as multilevel longitudinal models are sometimes called), have the added advantage over traditional repeated measures ANOVA of being flexible in terms of missing data and the coding of time (Bell et al, 2013). Specifically, while repeated measures ANOVA requires full information for all observed individuals and is limited to fixed time points, multilevel growth models are still functional in instances where some data are missing and can include time as a continuous predictor.

Why MLM rather than a structural equation/latent growth curve model? There are several plausible approaches to analyzing nested longitudinal data such as these. For example, both MLM and latent growth curve modeling (LGCM) could prove useful for modeling these data. While LGCM is drawn from a structural equation modeling (SEM) framework, MLM is an expansion of more common regression methods. While they are conceptually distinct in their approaches to model estimation, MLM and LGCM generally produce quantitatively comparable results (Bliese & Ployhart, 2002). Given its greater accessibility to audiences familiar with linear regression, MLM was therefore deemed the preferable approach.

Parameter estimation

While some have argued for the advantages of Bayesian estimation methods in fitting hierarchical linear models, maximum likelihood approaches are more commonly applied (Boedeker, 2017). This is likely due to the specialized training and software that is required for the application of Bayesian methods. In this case, maximum likelihood was selected as the most practicable approach. Maximum likelihood estimation entails: 1) the construction of a likelihood function, which is "an equation that expresses the probability of observing the sample data as a function of the model's unknown parameters (both the fixed and random effects)" and 2) a numerical examination of "the relative performance of competing alternative estimates of these unknown parameters until those values that maximize the likelihood function are found" (Singer, 2002, p. 157).

There are two common approaches to maximum likelihood estimation in the context of multilevel models: restricted maximum likelihood estimation (REML) and full maximum likelihood estimation (ML). REML has certain advantages over ML, including its conceptual appeal (REML does not entail the false conflation of estimated fixed effects with unobservable population parameters), as well as its more conservative approach to the estimation of degrees of freedom and random effects (REML adjusts for "uncertainty in the estimates of the fixed effects by adjusting the estimates of the random effects accordingly") (Singer, 2002, p. 158). In some instances, REML can therefore result in less biased parameter estimates than ML. ML, on the other hand, boasts the significant advantage of allowing for more useful comparison of model fit using standard fit statistics such as the -2 Log Likelihood (-2LL), Akaike Information Criterion

(AIC) and the Bayesian Information Criterion (BIC). Due to the partitioning of fixed and random components involved in REML, -2LL, AIC, and BIC statistics for models estimated using REML only compare the fit of random components (Bodeker, 2017; Singer, 2002).

On the other hand, -2LL, AIC, and BIC statistics for models estimated using ML compare fit of both random and fixed components, and can be used for both non-nested and nested model comparisons. Further, the risk of biased estimates that can be associated with ML is greatest in cases where the number of clusters is small, which is not the case here (Boedeker, 2017, Singer, 2002). According to Boedeker (2017), to assess the possibility of bias in ML, a model can be fit using each method (ML and REML) and the variance estimates compared. Large deviations between ML and REML estimates suggest the use of REML. This comparison was conducted for these data using unconditional means models with each method, and only minimal change in variance component estimates was observed (< 2%). Given the consistency of results between ML and REML, the large number of clusters, and the utility of being able to compare model fit for nested and non-nested models, ML estimation was employed for all remaining analyses. Denominator degrees of freedom were computed using the between/within method.

The unconditional means model

Following Singer (2002), my modeling process begins with fitting an unconditional means model, which specifies the nesting structure but includes no predictors, to assess overall variation in fidelity. As Singer states, “In this model, we do not explore any *systematic* variation in Y over time, instead simply quantifying the extent to which Y varies” (p. 144). One of the great utilities of a model such as this, however, is that it allows for the straightforward calculation of intraclass correlation coefficients (ICC) to assess the proportion of variation that is attributable to each level of nesting. Following the notation conventions of Snijders and Bosker (2012), the unconditional means model is expressed in Model 1.1 below:

Model 1.1 $$\mathbf{Y_{ijk.} = \beta_{0jk} + R_{ijk}}$$

Where $\mathbf{Y_{ijk}}$ is the fidelity score for session *i* conducted by provider *j* in organization *k.* $\mathbf{\beta_{0jk}}$ can be understood as the intercept for provider *j* within organization *k* (Snijders and Bosker, 2012). That intercept, in turn, can be expressed as the sum of an average intercept at the organization level, $\mathbf{\delta_{00k}}$, and an error term, $\mathbf{U_{0jk}}$:

$$\mathbf{\beta_{0jk} = \delta_{00k} + U_{0jk}}$$

$\mathbf{\delta_{00k}}$ itself contains a fixed component and a random variance component, and can be expressed as the following sum:

$$\mathbf{\delta_{00k} = \gamma_{000} + V_{00k}}$$

Model 1.1 can therefore be re-expressed in the following way:

Model 1.2 $$\mathbf{Y_{ijk.} = \gamma_{000} + V_{00k} + U_{0jk} + R_{ijk}}$$

It is possible to identify in Model 1.2 two different types of model parameter: fixed and random. It is crucial to remember that, while fixed effects (in this case, only $\mathbf{\gamma_{000}}$) are estimated directly, such estimates are *not* available for random effects. Instead, only variance components are estimated for random effects. Allowing the intercepts to vary randomly at both provider and organizational levels has yielded a model in which there are three sources of variability. Variation in fidelity attributable to clustering within providers is written as:

$$\mathbf{\tau^2_0 \text{ or } Var(U_{0jk})}$$

Variation in fidelity attributable to clustering within organizations is written as:

$$\mathbf{\varphi^2_0 \text{ or } Var(V_{00k})}$$

Finally, residual variance is written as:

$$\boldsymbol{\sigma^2} \textbf{ or Var}(\mathbf{R_{ijk}}).$$

And total variance is equal to

$$\boldsymbol{\sigma^2 + \varphi^2_0 + \tau^2_0}$$

Results of the unconditional means model predicting fidelity are reported in Table 7 below.

Table 7. Results of one 3-level unconditional means model predicting fidelity

Fixed Component	**β Estimate**	**S. E.**	**T-Statistic**	**P-Value**
γ_{000} = Fixed Intercept	94.11	0.23	404.55	<.0001*
Random Component	**τ Estimate**	**S. E.**	**Z-Statistic**	**P-Value**
Level three random component				
$\boldsymbol{\varphi^2_0}$ = Var(V_{00k}), Org Level Intercept Variance	4.84	0.98	4.93	<.0001*
Level two random component				
$\boldsymbol{\tau^2_0}$ = Var(U_{0jk}), Provider Level Intercept Variance	9.09	0.70	12.91	<.0001*
Level one variance				
$\boldsymbol{\sigma^2}$ = Var(R_{ijk}), Residual Variance	48.31	0.58	83.48	<.0001*
-2LL = 100479.5				
AIC = 100487.5				
BIC = 100500.1				

Interpreting the fixed component

The fixed intercept estimate, $\boldsymbol{\gamma_{000}}$ = 94.11, represents the average fidelity score for the average provider. In a study design where each participant had exactly the same number of observations, this value would be identical to overall mean fidelity in the sample. However, given that providers vary in the number of sessions for which they have fidelity scores, the fixed intercept estimate and mean fidelity in the overall sample differ slightly. This is due to the fact that the fixed intercept is calculated as a “mean of means” for each provider, not the mean of all individual observations (Singer, 2002).

Interpreting the random component

The provider level intercept variance, $\boldsymbol{\tau^2_0}$ or $\mathbf{Var(U_{0jk})}$, refers to the variance in fidelity between providers, and the organization level intercept variance $\boldsymbol{\varphi^2_0}$ or $\mathbf{Var(V_{00k})}$, refers to the variance in fidelity between organizations. In this case, the term "intercept" is used in place of "average" to facilitate the extension of the unconditional means model into the growth models presented below (Singer, 2002). Finally, $\boldsymbol{\sigma^2}$, or $\mathbf{Var(R_{ijk})}$, denotes variability in fidelity *within* individuals and organizations, known as residual variance.

Assessing the effects of clustering

It is possible to use a simple ratio to assess the proportion of variance accounted for by different levels of nesting. For example, we may compare the sum of the random intercepts (($\boldsymbol{\tau^2_0} + \boldsymbol{\varphi^2_0}$) = (4.84 + 9.09) = 13.93) with the residual variance ($\boldsymbol{\sigma^2}$ = 48.32) by dividing as follows:

$$\mathbf{48.32/13.93 = 3.47}$$

In this case, $\boldsymbol{\sigma^2}$ is nearly 3.5 times the size of $\boldsymbol{\tau^2_0} + \boldsymbol{\varphi^2_0}$, indicating that while providers and organizations do differ in their average fidelity scores, the majority of variation in fidelity occurs across sessions *within* providers and organizations (paraphrasing Singer, year, p. 149). A more formal measure of the effects of clustering in multilevel modeling is the ICC, which in a two-level growth model represents the "ratio of the variance component between persons and the variance components between and within persons" (Singer, year, p. 150). The two level version of the ICC can be straightforwardly extended for use in this three-level modeling context and can be calculated as follows:

$$\textbf{Organizational Level ICC} = \boldsymbol{\varphi^2_0/(\varphi^2_0 + \tau^2_0 + \sigma^2)} = \mathbf{4.84/(4.84 + 9.09 + 48.32) = .08}$$

$$\textbf{Individual Level ICC} = \boldsymbol{\tau^2_0/(\varphi^2_0 + \tau^2_0 + \sigma^2)} = \mathbf{9.09/(4.84 + 9.09 + 48.32) = .15}$$

The values calculated above provide useful information. We may now say that 8% of the total variability in fidelity is attributable to clustering within organizations, and that 15% of the total variability is attributable to clustering within providers. While the ICC for organization represents the total amount of variability uniquely stemming from that level of clustering, the likeness between providers within the same organization would be more accurately described by the sum of the provider level and organization level ICC's (8% + 15% = 28%). However, the majority of variability in fidelity (72%) is *not* explained by clustering at either level. In other words, most variability occurs *within* providers and *within* organizations, not *between* them. This raises the question of whether within provider and within organization variability can be understood as systematic change occurring over time.

The unconditional linear growth model

While the unconditional means model above is useful for assessing variability in fidelity overall, as well as the degree of similarity between sessions clustered within providers and organizations, it is not informative with regard to systematic change in fidelity over time. To assess this systematic change, it will be necessary to fit a growth model, beginning in this case with an unconditional linear growth model. In its simplest form, the growth model can be expressed as a session level regression model:

Model 2.1 $$\mathbf{Y_{ijk.} = \beta_{0jk} + \beta_{1jk}(Time)_{ijk} + R_{ijk}}$$

In the model above, "**Time**" represents the number of days since the first SafeCare session. In a similar fashion to Model 1.1, the beta coefficients in Model 2.1 can be expressed as the sum of a mean value and a random variance component:

$$\mathbf{\beta_{0jk} = \delta_{00k} + U_{0jk}}$$

$$\mathbf{\beta_{1jk} = \delta_{10k} + U_{1jk}}$$

And $\boldsymbol{\delta_{00k}}$ and $\boldsymbol{\delta_{10k}}$ can each be expressed as the sum of a fixed component and a random variance component:

$$\boldsymbol{\delta_{00k} = \gamma_{000} + V_{00k}}$$

$$\boldsymbol{\delta_{10k} = \gamma_{100} + V_{10k}}$$

So, through substitution, we can re-express Model 2.1 as follows:

Model 2.2.1 $\boldsymbol{Y_{ijk.} = \gamma_{000} + \gamma_{100}(Time)_{ijk} + U_{0jk} + U_{1jk}(Time)_{ijk} + V_{00k} + V_{10k}(Time)_{ijk} + R_{ijk}}$

The model statement above separates the fixed and random components while grouping the parameters by level of clustering. It could be equivalently expressed in a way that groups parameters related to the intercept, parameters related to the slope, and residual variance as follows:

Model 2.2.2 $\boldsymbol{Y_{ijk.} = [\gamma_{000} + U_{0jk} + V_{00k}] + [\gamma_{100}(Time)_{ijk} + U_{1jk}(Time)_{ijk} + V_{10k}(Time)_{ijk}] + R_{ijk}}$

It is important to note here that the inclusion of time as a covariate changes the interpretation of the fixed intercept term, $\boldsymbol{\gamma_{000}}$. Whereas $\boldsymbol{\gamma_{000}}$ represented a grand mean (or mean of means) in Model 1.2 (the unconditional means model), in Model 2.1 (the growth model) that parameter represents the mean value of fidelity *at baseline* (time = 0). This intercept is allowed to vary randomly at the provider level and the organization level, with the provider level variance component denoted as $\boldsymbol{U_{0jk}}$ and the organization level variance component denoted as $\boldsymbol{V_{00k}}$. So, the baseline score for provider *j* in organization *k* can be expressed as the sum of the average baseline score, $\boldsymbol{\gamma_{000}}$, and the provider level deviation from that average, $\boldsymbol{U_{0jk}}$, plus the organization level deviation from that average, $\boldsymbol{V_{0jk}}$.

Baseline fidelity for provider *i* in organization *k* $\boldsymbol{(\beta_{0jk}) = \gamma_{000} + U_{0jk} + V_{0jk}}$

Further, while Models 1.1 and 1.2 (the unconditional means model) assessed the variability in fidelity attributable to each level of nesting, as well as residual variance, by allowing the providers and organizations to vary randomly in terms of average fidelity (random intercepts), Models 2.1 and 2.2 allow for an additional type of random variation: namely, variation in the *relationship* between time and fidelity (random slopes). While in a traditional regression model, the relationship between time and fidelity would be estimated by the fixed effect coefficient, $\boldsymbol{\beta_1}$, that beta coefficient is allowed to vary in a random effects model such as Models 2.1 and 2.2 through the introduction of provider level and organization level error terms ($\mathbf{U_{1jk}}$ and $\mathbf{V_{10k}}$, respectively). Thus, for provider *j* within organization *k*, the effect (or slope) of time on fidelity can be expressed as the following sum:

Slope of time on fidelity for provider *j* in organization *k* $(\boldsymbol{\beta_{1jk}}) = \boldsymbol{\gamma_{100} + U_{1jk} + V_{10k}}$

where $\boldsymbol{\gamma_{100}}$ represents the average beta coefficient for time, $\mathbf{U_{1jk}}$ represents the provider level deviation from that average, and $\mathbf{V_{10k}}$ represents the organization level deviation. Again, the random intercepts and slopes at both the provider and organizational level are not directly estimated. Rather, the multilevel model estimates only their associated variance components. Following Snijders and Bosker (2012) they are denoted here as follows:

$$\textbf{Provider Level Intercept Variance} = \boldsymbol{\tau^2_0} = \mathbf{Var(U_{0jk})}$$

$$\textbf{Provider Level Slope Variance} = \boldsymbol{\tau^2_1} = \mathbf{Var(U_{1jk})}$$

$$\textbf{Organization Level Intercept Variance} = \boldsymbol{\varphi^2_0} = \mathbf{Var(V_{00k})}$$

$$\textbf{Organization Level Slope Variance} = \boldsymbol{\varphi^2_1} = \mathbf{Var(V_{10k})}$$

In addition to these four variances, this three-level random intercept and slope model (or growth model) estimates two more important parameters: the slope-intercept covariances. They can be expressed as follows:

$$\textbf{Provider Level Slope-Intercept Covariance} = \mathbf{Cov(U_{0jk}, U_{1jk})} = \boldsymbol{\tau_{01}}$$

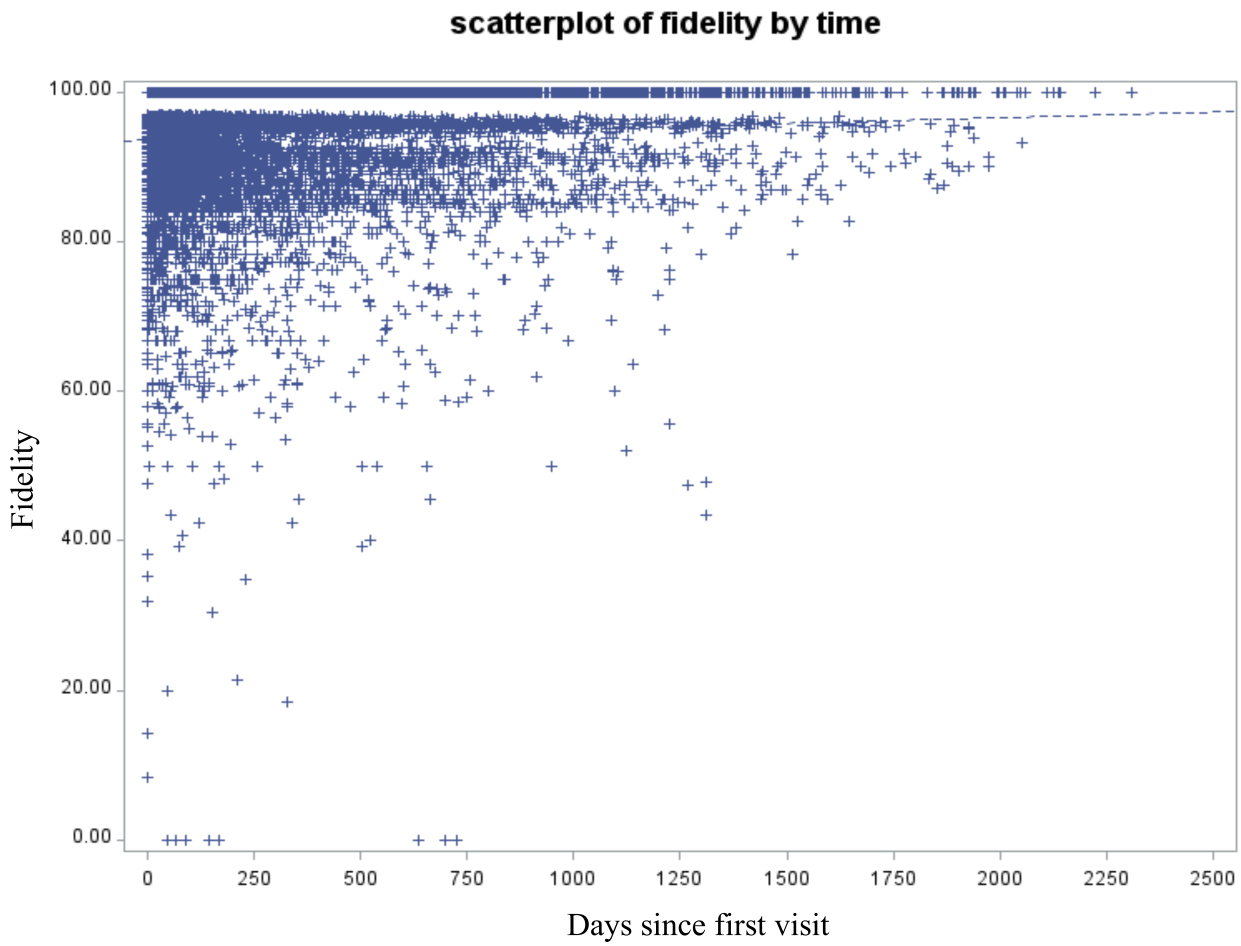
$$\textbf{Organization Level Slope-Intercept Covariance} = \mathbf{Cov(V_{00k}, V_{10k})} = \boldsymbol{\varphi_{01}}$$

The covariance parameters are a function of the correlation between random intercepts and slopes, and can provide useful information regarding the relationship between baseline values and trends over time.

Functional form and model assumptions

While there is a significant degree of heterogeneity in the methodological literature on longitudinal/multilevel models such as these, there is broad agreement across sources regarding the importance of identifying the appropriate functional form of the relationship between time and the outcome of interest (Snijders & Bosker, 2012; Singer, 2002; Bliese & Ployhart, 2002). More concretely, the growth model outlined above assumes that the relationship between time and fidelity can be reasonably approximated by a linear function. This assumption cannot be taken for granted, however, and further exploration is needed to determine whether a different function would fit the data better. Bliese and Ployhart (2002) suggest that this exploration should occur prior to the inclusion of any covariates in addition to time. Consequently, determining the appropriate functional form of the relationship between predictors and outcome is an analytic step that requires considerable attention at this stage. My first step is a visual inspection of scatterplots for the distribution of fidelity by time overall (see Figure 5 below). This inspection suggests that fidelity is generally high and stable over time (although there is clearly a larger range of scores at baseline and in early observations than in later observations, with a small number of very low scores near baseline). There is also some indication of a possible ceiling effect, with many sessions receiving the full score of 100 (an indication that is supported by the univariate distribution of fidelity score shown in Chapter 3.1.1, Figure 4).

Figure 5. Scatterplot of fidelity by time including OLS regression line



The assessment of the linearity assumption is significantly more complicated in the case of longitudinal multilevel models than in standard regression models, however, and even more so in cases where levels two and three have a large number of clusters (in this case 868 providers and 172 organizations), since the assumption of linearity must be assessed between predictor and outcome within *each* cluster. Visual inspection of a single scatterplot including each provider's growth curve to assess the tenability of a linear relationship between time and fidelity within providers does not yield much insight, given the overwhelming amount of information presented. The same challenge arises in a similar plot including each organization's growth curve (See Figures 6 and 7 below).

Figure 6. Scatterplot of fidelity by time with growth curves for each provider

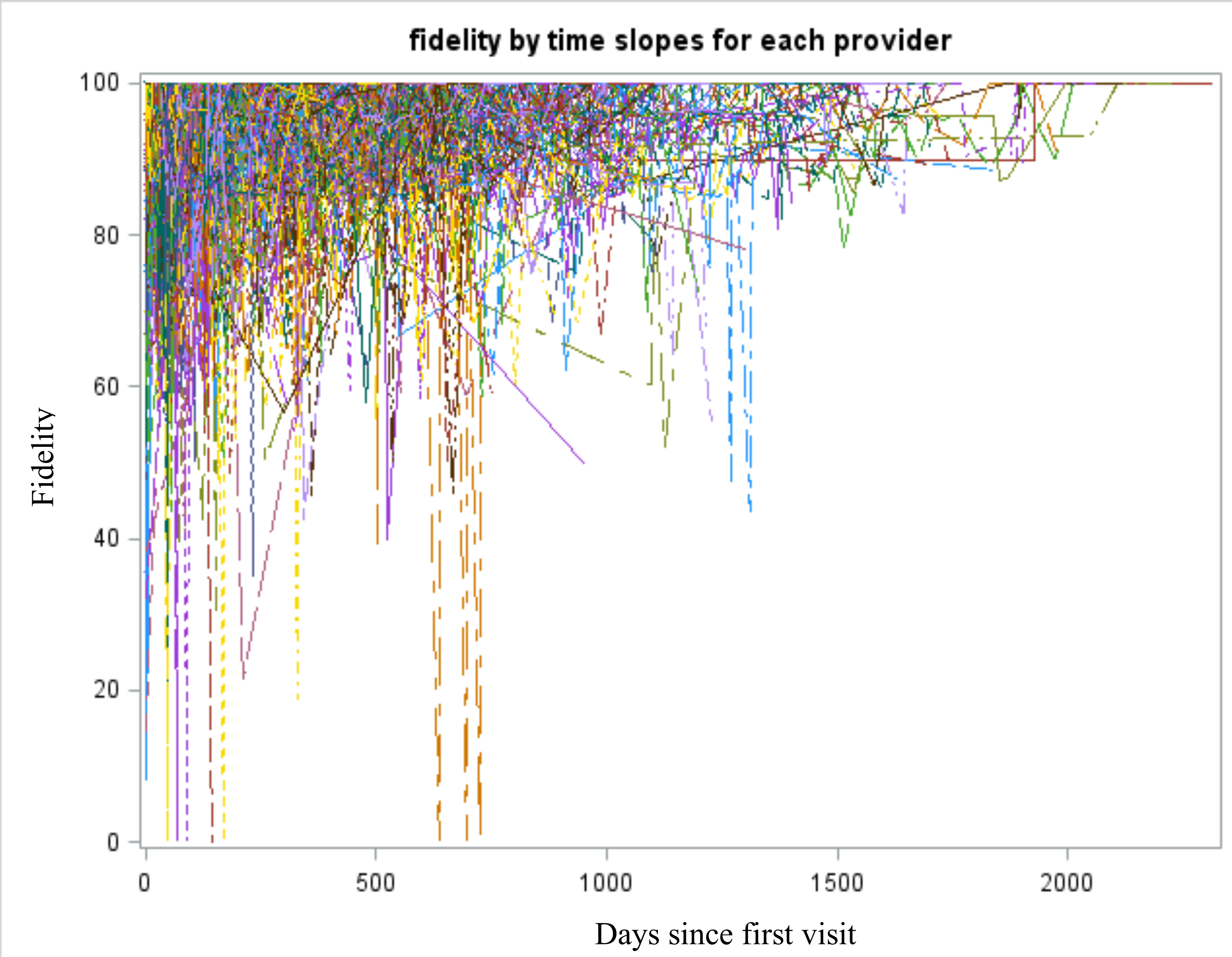
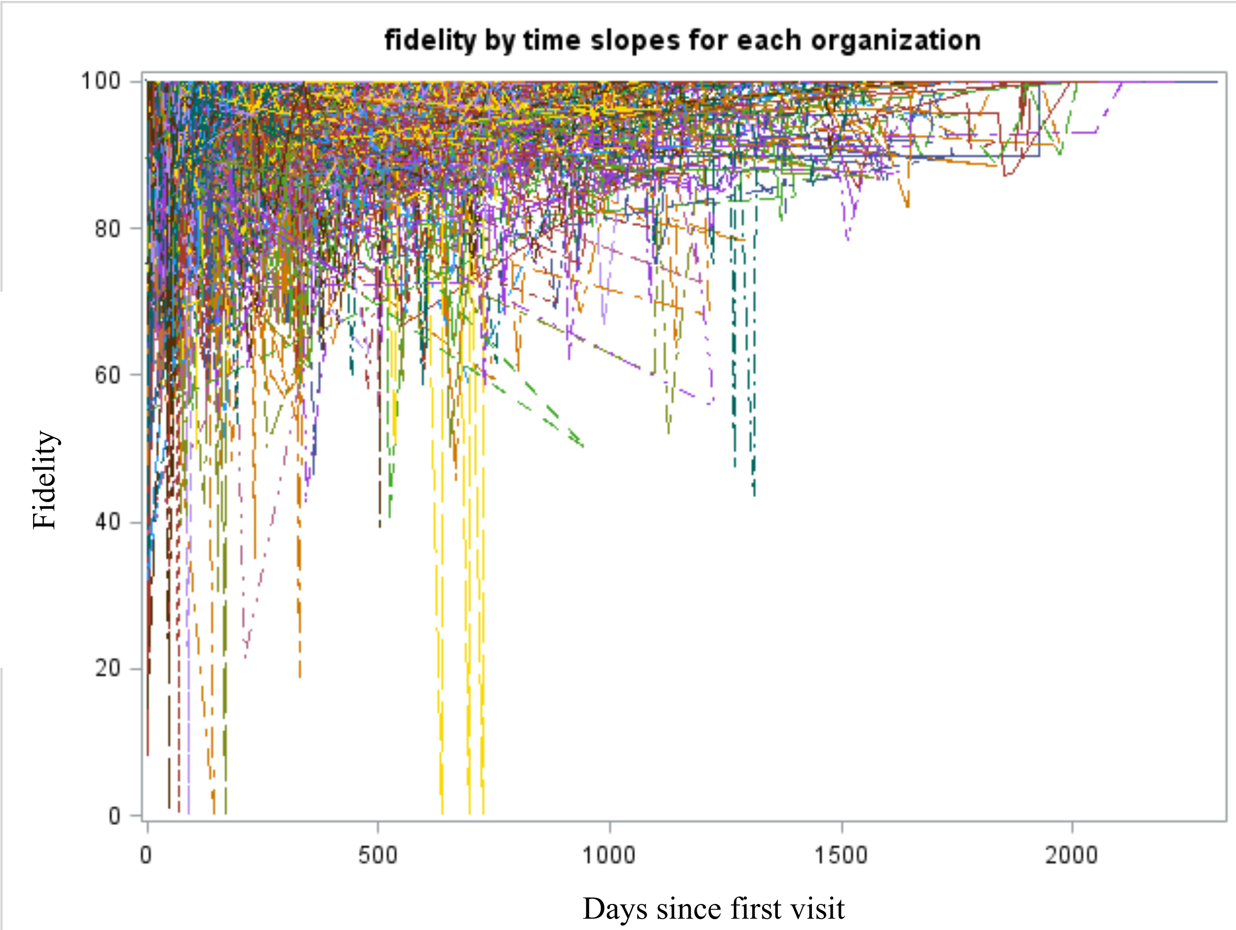


Figure 7. Scatterplot of fidelity by time with growth curves for each organization



Given the impracticality of assessing functional form using an overall scatterplot for providers and an overall scatterplot for organizations, individual scatterplots of fidelity by time were generated and consulted for all 868 providers and all 172 organizations in the sample. This process revealed a significant degree of heterogeneity in the form of this relationship, and no consistent pattern was apparent.

In the absence of a clear visual indicator of functional form, several quantitative approaches to identifying the appropriate form for inferential models were considered, including fitting a model containing a quadratic term ($time^2$). Bliese and Ployhart (2002) recommend an iterative approach to model building that includes the assessment of a linear relationship between time and the outcome of interest, followed by quadratic and higher order polynomial functions. After fitting a growth model including time as a fixed effect and a random effect at both provider and organization levels, I fit another model which included an added fixed effect, time^2. While Bliese and Ployhart (2002) recommend using statistical significance tests to determine inclusion or exclusion of quadratic terms, that approach was amended here for several reasons. The statistical hypothesis test for time^2 rejected the null hypothesis with a T-value approaching infinity, and the effect size was infinitesimal ($B = 2.25 \times 10^{-8}$, or .0000000225, $p < .0001$). The interest of this project is not in fidelity over time across the entire sample, however, but in fidelity trajectories *within providers.* Addressing a possible violation of the linearity assumption could therefore not be achieved using a fixed quadratic term that did not vary between providers. Thus, my next step was to fit another model which allowed the time^2 parameter to vary at the provider level. At this point, computational issues arose and the algorithm failed to converge.

Snijders and Bosker (2012) note that computational issues can arise when the intercept for time is at an extreme value in the distribution (such as 0, or baseline). Therefore, time was recoded to be centered at its mean, and the new mean-centered time variable (along with an associated quadratic term) were included in the model with a fixed effect for time (varying randomly at provider and organizational levels) and for time^2 (varying randomly at the provider level). However, computational issues arose again. Given the difficulties in utilizing the quadratic term to model the relationships of interest (fidelity trajectories *within* providers and differences in those trajectories *between* providers), I reconsidered the inclusion of the fixed quadratic term altogether. Consulting the fit statistics for the growth models including and excluding the quadratic term for time, I found that the fit of the model was better without time^2. While the deviance statistic (-2 Log Likelihood) remained unchanged between the two models

(97871.2), the AIC and BIC statistics were lower for the model without the quadratic term (AIC decreased from 97891.2 to 97889.2, and BIC decreased from 97922.7 to 97917.5). Therefore, the quadratic function was abandoned in favor of the initial linear function.

While the linearity assumption is of primary importance, both homoscedasticity and normality of residuals were considered as well. Other transformations to improve model fit were explored, including logarithmic and square root transformations of time (and other continuous predictors in later models), as well as transformations on the fidelity outcome. While it is possible that they may have improved model fit to some extent, these approaches were not ultimately selected for several reasons. First, as Snijders and Bosker (2012) state with regard to the linearity assumption, even in cases where the fit of the model's functional form is only moderate, multilevel models can still provide a reasonable approximation of overall longitudinal trends on average. Second, interpretability of model parameters was considered highly important given the real-world nature of the data. Variable transformations would inevitably increase the interpretive complexity of the models fit, and this complexity was weighed against the possible benefit of a moderate improvement in model fit. Third, as will be discussed below, comparing the residual variances, $\boldsymbol{\sigma}^2$, of the unconditional means model and the unconditional linear growth model (where the latter differs from the former only by the inclusion of time as a fixed and random effect), I was able to determine that only a small proportion of the original within-person variability in fidelity was attributable to time. Other predictors in later models also accounted for only a small portion of variability. Given the low overall variability in fidelity and the small proportion of explained variability attributable to time and other predictors, the interpretive complexity added by including logarithmic or other transformations was deemed not worthwhile. Further, since the goal of Research Question 1 is to develop a general understanding of whether and how fidelity tends to change over time, testing the linear relationship of fidelity with an untransformed time variable was ultimately selected as the optimal approach. The results of the unconditional linear growth model are reported in Table 8 below.

Table 8. Results of one 3-level unconditional linear growth model predicting fidelity

Fixed Component	**β Estimate**	**S. E.**	**T-Value**	**P-Value**
γ_{000} = Fixed Intercept	93.85	0.25	380.27	<.0001*
γ_{100} = Fixed Slope for Time	0.001	0.0004	2.99	0.0028*
Random Component	**τ Estimate**	**S. E.**	**Test Statistic**	**P-Value**
Level three random component				
φ^2_0 = Var(V_{00k}), Org Level Intercept Variance	4.87	1.12	4.34	<.0001*
φ^2_1 = Var(V_{10k}), Org Level Slope Variance	2.696E-6	3.809E-6	0.71	0.2396
φ_{02} = Cov(V_{00k}, V_{10k}), Org Level Intercept-Slope Covariance	0.0005	0.001	0.35	0.7298
Level two random component				
τ^2_0 = Var(U_{0jk}) Provider Level Intercept Variance	11.71	1.00	11.70	<.0001*
τ^2_1 = Var(U_{1jk}) Provider Level Slope Variance	0.00003	6.18E-6	4.22	<.0001*
τ_{02} = Cov(U_{00k}, U_{10k}), Org Level Intercept-Slope Covariance	-0.01	0.002	-4.64	<.0001*
Level one variance				
σ^2 = Var(R_{ijk}), Residual Variance	45.69	0.57	80.80	<.0001*
-2LL = 9781.2 AIC = 97889.2 BIC = 97917.5				

Interpreting the fixed component

The value of 93.85 for the fixed intercept, γ_{000}, represents the average baseline fidelity. Similarly, the value of γ_{100} = 0.001 represents the average slope, or growth rate. So, we may say that the average provider had a score of 93.85% fidelity at baseline, with each passing day being statistically significantly associated with an increase of 0.001% in model predicted fidelity score ($p = .0028$). In Singer's (2002) words, these results indicate that "on average, there *is* systematic linear change over time" (p. 154). While this growth rate is very small, it is worth noting that the median number of days since first visit was 141, the mean was 288, and the maximum was 2311. This means that model predicted fidelity increased by 0.141 percentage points between baseline and the median number of days, by 0.288 percentage points between baseline and the mean number of days, and by 2.311 percentage points between baseline and the maximum number of days observed in the sample. In other words, these findings provide further evidence for the indications in the scatterplot of fidelity by time: that fidelity is high, fairly stable, and tends to increase slightly over time.

Interpreting the random component

While the fixed component of the model provides useful information about what is occurring on average, the random component can help us to understand differences between providers and between organizations. The organization level intercept variance, $\boldsymbol{\varphi^2_0} = 4.87$, for example, describes the variability in baseline fidelity that is accounted for by clustering within organizations, and the associated null hypothesis test indicates that this variability is statistically significantly different from 0 ($p < .0001$). Similarly, the provider level intercept variance, $\boldsymbol{\tau^2_0} = 11.71$, describes the variability in baseline fidelity that is accounted for by clustering within providers, and the associated null hypothesis test indicates that this variability is also statistically significantly different from 0 ($p < .0001$). Due to the inclusion of time as a predictor in the model, these intercept variances cannot be simply compared to the residual variance, $\boldsymbol{\sigma^2}$, as was done above using the ICC's for the unconditional means model (Model 1). We can, however, compare them to one another using a ratio:

$$\mathbf{11.71 / 4.87 = 2.40}$$

This ratio indicates that clustering within providers explains 2.4 times as much variability in baseline fidelity as does clustering within organizations, or similarly that there is 2.4 times as much between-provider variability in baseline fidelity as there is between-organization variability.

The organization level slope variance, $\boldsymbol{\varphi^2_1}$ = 2.696 x 10^{-6}, represents the amount of variability in growth over time that is attributable to clustering at the organizational level. Unsurprisingly, this miniscule amount of variability is not statistically significantly different from 0 ($p = .2396$), and accordingly we must say that we have failed to detect any difference in growth rate attributable to clustering within organizations. The somewhat larger provider level slope variance, $\boldsymbol{\tau^2_1}$ = 0.000026, is statistically significantly different from 0 ($p < .0001$). Given the already small amount of change in fidelity over time, however, this amount of variability may arguably be practically non-significant.

The covariance parameter for organization level intercept and organization level slope, $\boldsymbol{\varphi_{02}}$ = .001, is not statistically significant ($p = .7298$). This indicates that there is no relationship between the organization level contribution to baseline fidelity and the organizational level contribution fidelity over time. However, the corresponding provider level covariance parameter, $\boldsymbol{\tau_{02}}$ = -.01, was statistically significant ($p < .0001$). The negative value for this parameter indicates that providers who have higher baseline values of fidelity tend to have slightly smaller slopes on fidelity over time (their growth rates are slightly lower than the growth rates for providers with lower baseline fidelity). This may be due to a ceiling effect, given that fidelity tends to be high overall.

The residual variance, $\boldsymbol{\sigma^2}$ = 45.69, is the within person variance component. Unlike the other random effects interpreted above, it *does* maintain the same interpretation between the unconditional means and unconditional linear growth models. Following Singer (2002), we may assess the proportion of variance explained by time by subtracting the $\boldsymbol{\sigma^2}$ value for the growth model from the $\boldsymbol{\sigma^2}$ value from the means model, and then dividing that difference by the means model $\boldsymbol{\sigma^2}$ as follows:

$$\mathbf{48.32 - 45.69 = 2.63}$$

$$\mathbf{2.63/48.32 = .05}$$

This value indicates that approximately 5% of the "original within person variability is 'explained by time'" (Singer, 2002).

Interpreting model fit statistics

Each of the three key model fit statistics (-2LL, AIC, and BIC) favored the growth model over the means model, indicating that, despite the small effect size, the inclusion of time as a predictor improved the model fit.

Conditional linear growth models with cross level direct effects

In this section, I report the results of a series of linear growth models that are identical to Model 2 above, except for the additional inclusion of one predictor variable each. These models serve a similar function to that which a series of bivariate models would serve in iterating a model taxonomy to produce a final multivariable model in a standard linear regression framework. According to Aguinis et al. (2013), there are three types of effect that can be tested in an MLM framework: lower-level direct effects, cross-level direct effects, and cross-level interaction effects. In the case of the growth model fit above, one lower-level direct effect was tested: the relationship between session fidelity (a session level outcome variable) and time (a session level predictor), to assess the overall linear trajectory of fidelity over time. In addition to its fixed effect, time was also allowed to vary randomly at the provider and organization levels to account for possible differences in fidelity trajectory between providers and organizations. In the models reported in this section, each linear growth model includes an additional cross-level direct effect, with a provider level variable predicting session level fidelity score. At this juncture, the models do not provide any information about the relationship between the additional predictor and fidelity *trajectories*, but rather they characterize the relationship between that predictor and fidelity overall. Predictor effects on fidelity trajectories will be assessed in the section reporting conditional linear growth models with cross level direct and interaction effects below.

I considered the inclusion of aggregate variables at the organization level (generated by aggregating the information available from the providers nested therein) as predictors during this step, but given the low proportion of variability in fidelity explained at the organization level (as indicated by the ICC value for that level of nesting in Model 1, 8%), I determined that the inclusion of these variables was not worth the associated increase in model complexity. This decision was further supported by the fact that my central theoretical interest was in the relationship between individual level variables and fidelity. The models fit in this section can be expressed similarly to the growth models fit above, with the addition of a single predictor at the provider level. For continuous predictors, they can be expressed as follows:

Model 3.1 $\mathbf{Y_{ijk.} = \beta_{0jk} + \beta_{1jk}(Time)_{ijk} + \beta_{2jk}(Predictor)_{jk} + R_{ijk}}$

For categorical predictors with >2 levels, the model expression must be amended to account for dummy coding (as produced through a CLASS statement in SAS PROC MIXED) as follows:

Model 3.2

$$\mathbf{Y_{ijk.} = \beta_{0jk} + \beta_{1jk}(Time)_{ijk} + \beta_{2jk}(Predictor\ 2\ vs.\ 0)_{jk} + \beta_{3jk}(Predictor\ 1\ vs.\ 0)_{jk} + R_{ijk}}$$

In this model, "Predictor" stands in for each of the specific predictors reported in Table 9 below, and the intercept, $\boldsymbol{\beta_{0jk}}$, and slope for time, $\boldsymbol{\beta_{0jk}}$, can be decomposed into their fixed and random components as in Model 2.1 (the unconditional linear growth model). As a fixed effect, $\boldsymbol{\beta_{2jk}}$ (or $\boldsymbol{\beta_{2jk}}$ and $\boldsymbol{\beta_{3jk}}$ for categorical predictors with >2 levels), has no random component. In Table 9 below, the central focus is on the magnitude and statistical significance of effects of predictor variables. For categorical predictor variables with more than 2 levels, F-tests (type 3 tests of fixed effects) were reported alongside T-tests in order to assess overall significance. In the interest of space and concision, variance components were not reported for these models. Time was also included in each of the models reported in Table 9, but those parameter estimates were excluded from the table due to negligible changes from one model to another.

Table 9. Results of several 3-level linear growth models each including a single predictor

Fixed Effect	**(γ_{000}) Intercept**	**β Estimate**	**S. E.**	**T-Value**	**P-Value**
EBPAS Positivity					
$\boldsymbol{\beta_{2jk}}$ = EBPAS Positivity	93.89	0.84	0.24	3.54	0.0004*
EBPAS Negativity					
$\boldsymbol{\beta_{2jk}}$ = EBPAS Negativity	93.89	-0.66	0.21	-3.21	0.0013*
Provider Gender					
$\boldsymbol{\beta_{2jk}}$ = Female	93.29	0.60	0.54	1.10	0.2768
Male		Ref			
Provider Race					
$\boldsymbol{\beta_{2jk}}$ = Non-White	94.06	-0.48	0.31	-1.55	0.1261
White		Ref			

Provider Education					
β_{2jk} = Graduate Degree	93.32	1.16	0.46	2.54	0.0124*
β_{3jk} = Bachelor's Degree		0.36	0.40	0.92	0.3616
High School		Ref			
Type 3 Test of Fixed Effects				F = 3.92	0.0223*
Full Time Status					
β_{2jk} = Part Time	93.86	-0.09	0.42	-0.22	0.8282
Full Time		Ref			
Years on the Job					
β_{2jk} = Years on the Job	93.89	-0.01	0.03	-0.41	0.6817
Experience Serving at Risk Families					
β_{2jk} = (2) >1 Year	92.60	1.41	0.57	2.46	0.0156*
β_{3jk} = (1) <1 Year		1.33	0.49	2.72	0.0078*
(0) None		Ref			
Type 3 Test of Fixed Effects				F = 3.90	0.0237*
Exp. Serving Substantiated Families					
β_{2jk} = (2) >1 Year	93.16	0.77	0.43	1.79	0.0758
β_{3jk} = (1) <1 Year		0.80	0.49	1.62	0.1083
(0) None		Ref			
Type 3 Test of Fixed Effects				F = 1.72	0.1839
Exp Training Providers w/ high risk families					
β_{2jk} = (2) >1 Year	93.92	-0.46	0.32	-1.44	0.1520
β_{3jk} = (1) <1 Year		0.42	0.46	0.92	0.3583
(0) None		Ref			
Type 3 Test of Fixed Effects				F = 1.79	0.1703
Exp. providing Struc. Parenting					
β_{2jk} = (2) >1 Year	93.64	0.15	0.32	0.47	0.6409
β_{3jk} = (1) <1 Year		0.61	0.37	1.65	0.1019
(0) None		Ref			
Type 3 Test of Fixed Effects				F = 1.44	0.2392

Exp. Train. Providers in Struc. Parenting					
β_{2jk} = (2) >1 Year	94.04	-0.93	0.34	-2.73	0.0072*
β_{3jk} = (1) <1 Year		0.06	0.42	0.14	0.8853
(0) None		Ref			
Type 3 Test of Fixed Effects				F = 3.98	0.0208*
Learned Structured Parenting					
β_{2jk} = Yes	93.66	0.39	0.27	1.41	0.1615
No		Ref			
Prior Training EBI					
β_{2jk} = Yes	93.74	0.30	0.29	1.05	0.2954
No		Ref			

Interpreting the cross level direct effects

Both of the mean-centered EBPAS factor variables (positivity and negativity) were statistically significantly associated with fidelity. The directions of these associations were consistent with those hypothesized. Each one point increase in EBPAS positivity was associated with fidelity scores that were 0.84 percentage points higher on average (p = .0004), while each one point increase in EBPAS negativity was associated with fidelity scores that were 0.66 percentage points lower on average (p = .0013). Overall, provider education was associated with fidelity scores (F = 3.92, p = .0223). Providers with graduate degrees had fidelity scores that were on average 1.16 percentage points higher than providers with only high school diplomas (p = .0124). While providers with bachelor's degrees had fidelity scores that were 0.36 percentage points higher on average than providers with only high school diplomas, that difference was not statistically significant (p = .3616). Provider experience serving at risk families was associated with fidelity as well (F = 3.90, p = .0237). Providers who had greater than one year of experience serving at risk families had fidelity scores that were 1.41 percentage points higher on average than providers with no experience (p = .0156). Providers with some, but less than one year, of experience serving at risk families had fidelity scores that were 1.3 percentage points higher on average than providers with no experience ($p = .0078$). Experience training *other* providers in structured parenting interventions was statistically significantly associated with fidelity ($F = 3.98$, $p = .0208$). Providers with greater than one year of experience training other providers in structured parenting interventions had, on average, fidelity scores that were 0.93 percentage points lower than providers with no such experience ($p = .0072$). Between providers with some, but less than one year, of experience training other providers in structured parenting interventions and providers with no such experience, there was no detectable difference in fidelity on average.

Conditional linear growth models with cross level direct and interaction effects

The models reported below are similar to the conditional growth models with cross level direct effects with one distinction: in addition to cross level direct effects, cross-level interactions were also tested to determine whether the relationship between time and session fidelity was impacted by differences in provider level predictor variables. For continuous and dichotomous predictors, these models can be expressed as follows:

Model 4.1 $\mathbf{Y_{ijk.} = \beta_{0jk} + \beta_{1jk}(Time)_{ijk} + \beta_{2jk}(Predictor)_{jk} + \beta_{3jk}(Time_{ijk} \text{ x } Predictor_{jk}) + R_{ijk}}$

For categorical predictors with >2 levels, the model expression must be amended to account for dummy coding as follows:

Model 4.2

$$\mathbf{Y_{ijk.} = \beta_{0jk} + \beta_{1jk}(Time)_{ijk} + \beta_{2jk}(Predictor\ 2\ vs.\ 0)_{jk} + \beta_{3jk}(Time_{ijk} \text{ x } Predictor\ 2\ vs.\ 0)_{jk} + \beta_{4jk}(Predictor\ 1\ vs.\ 0)_{jk} + \beta_{5jk}(Time_{ijk} \text{ x } Predictor\ 1\ vs.\ 0)_{jk} + R_{ijk}}$$

Much like before, the intercept, $\boldsymbol{\beta_{0jk}}$, and the slope for time, $\boldsymbol{\beta_{1jk}}$, can be decomposed into their fixed and random components. The inclusion of the time by predictor interaction term results in a change in the interpretation of the beta coefficients, however. For the coefficient on time, $\boldsymbol{\beta_{1jk}}$, the interpretation remains much the same, save for the addition of "controlling for the covariate," (Singer, 2002). For the direct effect of the predictor, $\boldsymbol{\beta_{2jk}}$, the beta coefficient no longer refers to the relationship between that variable and fidelity on average overall, but rather between that variable and *initial* fidelity (Singer, 2002). In other words, this coefficient refers to the effect of that covariate on fidelity *when the interaction term = 0.* The coefficient on the interaction term, $\boldsymbol{\beta_{3jk}}$, in turn, refers to the difference in growth rate associated with a one unit difference in continuous predictor or a categorical difference in categorical predictors.

Table 10. Results of 3-level models predicting fidelity over time, including time and time by predictor interaction terms

Predictor	Intercept	β Estimates	S. E.	T-Value	P-Value
EBPAS Positivity					
β_{2jk} = EBPAS Positivity	93.89	0.82	0.27	3.01	0.0026*
β_{3jk} = Positivity*Time		0.0001	0.0006	0.15	0.8801
EBPAS Negativity					
β_{2jk} = EBPAS Negativity	93.89	-0.69	0.23	-2.94	0.0032*
β_{3jk} = Negativity*Time		0.0002	0.0006	0.28	0.7810
Provider Gender					
β_{2jk} = Female	93.12	0.78	0.62	1.25	0.2200
β_{3jk} = Female*Time		-0.0009	0.0016	-0.58	0.5597
Male		Ref			
Provider Race					
β_{2jk} = Non-White	94.04	-0.43	0.35	-1.24	0.2185
β_{3jk} = Non-White*Time		-0.0003	0.00082	-0.36	0.7224
White		Ref			
Provider Fulltime					
β_{2jk} = Part Time	93.81	0.26	0.48	0.55	0.5873
β_{3jk} = Part Time*Time		-0.002	0.001	-1.59	0.1109
Full Time		Ref			

Years on the Job					
β_{2jk} = Years on the Job	93.86	-0.003	0.04	-0.07	0.9403
β_{3jk} = Years OTJ*Time		-0.00006	0.000088	-0.66	0.5105
Provider Education					
β_{2jk} = Graduate Degree	93.09	1.51	0.51	2.95	0.0039*
β_{3jk} = Grad*Time		-0.002	0.001	-1.47	0.1411
β_{4jk} = Bachelor's Degree		0.63	0.45	1.41	0.1625
β_{5jk} = Bachelor's * Time		-0.001	0.001	-1.30	0.1928
High School		Ref			
Type 3 Test (Predictor)				F = 4.71	0.0107*
Type 3 Test (Interaction)				F = 1.21	0.2970
Exp. Serv. at Risk Fam.					
β_{2jk} = (2) >1 year	92.72	1.26	0.57	2.21	0.0293*
β_{3jk} = (2) >1 year*Time		0.0003	0.001	0.25	0.8059
β_{4jk} = (1) <1 year		0.95	0.66	1.42	0.1577
β_{5jk} = (1) <1 year*Time		0.002	0.002	1.44	0.1487
(0) None		Ref			
Type 3 Test (Predictor)				F = 2.53	0.0850
Type 3 Test (Interaction)				F = 1.71	0.1807
Exp. Serv Subst. Families					
β_{2jk} = (2) >1 year	93.30	0.65	0.49	1.33	0.1867
β_{3jk} = (2) >1 year*Time		0.0005	0.001	0.47	0.6384
β_{4jk} = (1) <1 year		0.42	0.57	0.75	0.4564
β_{5jk} = (1) <1 year*Time		0.002	0.001	1.38	0.1679
(0) None		Ref			
Type 3 Test (Predictor)				F = 0.93	0.3981
Type 3 Test (Interaction)				F = 1.16	0.3150
Train prov - HR fam.					
β_{2jk} = (2) >1 year	93.83	-0.14	0.37	-0.40	0.6896
β_{3jk} = (2) >1 year*Time		-0.002	0.01	-1.73	0.0835
β_{4jk} = (1) <1 year		0.52	0.52	0.99	0.3229
β_{5jk} = (1) <1 year*Time		-0.0005	0.001	-0.36	0.7203
(0) None		Ref			
Type 3 Test (Predictor)				F = 0.68	0.5087
Type 3 Test (Interaction)				F = 1.50	0.2234

Exp – Structured Par.					
β_{2jk} = (2) >1 year	93.75	0.15	0.37	0.40	0.6931
β_{3jk} = (2) >1 year*Time		0.00003	0.0009	0.03	0.9727
β_{4jk} = (1) <1 year		0.09	0.43	0.21	0.8339
β_{5jk} = (1) <1 year*Time		0.003	0.001	2.59	0.0095*
(0) None		Ref			
Type 3 Test (Predictor)				F = 0.08	0.9248
Type 3 Test (Interaction)				F = 4.04	0.0176*
Train prov- Struc. Par.					
β_{2jk} = (2) >1 year	94.05	-1.00	0.39	-2.53	0.0125*
β_{3jk} = (2) >1 year*Time		0.0003	0.001	0.31	0.7532
β_{4jk} = (1) <1 year		0.056	0.48	0.12	0.9067
β_{5jk} = (1) <1 year*Time		0.00002	0.001	0.02	0.9865
(0) None		Ref			
Type 3 Test (Predictor)				F = 3.40	0.0360*
Type 3 Test (Interaction)				F = 0.05	0.9509
Learned Struc. Parenting					
β_{2jk} = Yes	93.51	0.74	0.31	2.38	0.0194*
β_{2jk} = Yes*Time		-0.002	0.0008	-2.43	0.0150*
No		Ref			
Prior Training EBI					
β_{2jk} = Yes	93.85	-0.01	0.33	-0.04	0.9684
β_{2jk} = Yes*Time		0.002	0.0008	2.01	0.0439*
No		Ref			

Interpreting the cross level direct and interaction effects

Neither EBPAS factor was significantly associated with fidelity growth rates. Growth was associated, however, with prior experience implementing a structured parenting intervention, with providers who have some, but less than one year of experience displaying fidelity growth that is .003 percentage points higher than providers with no such experience (p = .0095). However, having greater than one year of experience was not associated with differences in fidelity growth. Fidelity growth was also associated with having learned structured parenting in the past, with those who answered yes having fidelity slopes that were, on average, .002%

smaller than those who answered no (p = .0150). Providers who had received prior training in evidence-based interventions had fidelity slopes that were, on average, .002% larger than those who had not (p = .0439)

Model 5: the final multivariable growth model

The final model is presented below in Table 11. This model contains fixed and random intercepts, a fixed and random slope for time, cross level direct effects for all variables that were statistically significant, and the cross level interaction effects that were statistically significant in previous analyses. While the model intercept and the slope for time each have fixed and random components, the beta coefficients for the remaining variables are fixed effects.

Table 11 Results of the final multivariable 3-level model predicting fidelity over time

Fixed Component	**β Estimate**	**S. E.**	**T-Value**	**P-Value**
Intercept				
γ_{000} = Fixed Intercept	92.47	0.59	156.39	<.0001*
Time in Days				
γ_{100} = Fixed Slope for Time	0.001	0.0007	2.20	0.0280*
EBPAS Positivity				
β_{2jk} = EBPAS Positivity	0.73	0.24	3.10	0.0019*
EBPAS Negativity				
β_{3jk} = EBPAS Negativity	-0.51	0.20	-2.51	0.0121*
Provider Education				
β_{4jk} = (3) Graduate Degree	0.87	0.46	1.88	0.0623
β_{5jk} = (2) Bachelor's Degree	-0.006	0.39	-0.02	0.9873
(1) High School	Ref			
Exp. Serv. AR Families				
β_{6jk} = (2) >1 Year	1.29	0.52	2.46	0.0160*
β_{7jk} = (1) <1 Year	1.30	0.60	2.16	0.0338*
(0) None	Ref			

Train prov- Struc. Par.				
β_{8jk} = (2) >1 year	-1.22	0.41	-3.01	0.0031*
β_{9jk} = (1) <1 year	-0.21	0.44	-0.47	0.6416
(0) None	Ref			
Exp – Structured Par.				
β_{10jk} = (2) >1 year	0.48	0.43	1.10	0.2712
β_{11jk} = (2) >1 year*Time	-0.0008	0.0009	-0.98	0.3264
β_{12jk} = (1) <1 year	-0.19	0.46	-0.42	0.6731
β_{13jk} = (1) <1 year*Time	0.002	0.001	2.25	0.0243*
(0) None	Ref			
Learned Struct. Par.				
β_{14jk} = Yes	0.59	0.35	1.69	0.0938
β_{15jk} = Yes*Time	-0.003	0.0008	-3.36	0.0008*
No				
Prior Training EBI				
β_{16jk} = Yes	-0.54	0.37	-1.48	0.1417
β_{17jk} = Yes*Time	0.003	0.0009	3.26	0.0011*
No				

Random Component	**τ Estimate**	**S. E.**	**Test Statistic**	**P-Value**
Level three random component				
φ^2_0 = Var(V_{00k}), Org Level Intercept Variance	4.54	1.06	4.27	<.0001*
φ^2_1 = Var(V_{10k}), Org Level Slope Variance	0.001	0.001	0.68	0.4934
φ_{02} = Cov(V_{00k}, V_{10k}), Org Level Intercept-Slope Covariance	8.233E-7	3.802E-6	0.22	0.4143
Level two random component				
τ^2_0 = Var(U_{0jk}) Provider Level Intercept Variance	10.78	0.97	11.13	<.0001*
τ^2_1 = Var(U_{1jk}) Provider Level Slope Variance	-0.009	0.002	-4.51	<.0001*
τ_{02} = Cov(U_{00k}, U_{10k}), Provider Level Intercept-Slope Covariance	0.00002	5.857E-6	3.86	<.0001*
Level one variance				

$\boldsymbol{\sigma^2}$ = Var(R_{ijk}), Residual Variance	45.50	0.57	80.36	<.0001*
-2LL = 96643.3 AIC = 96693.3 BIC = 96772.0				

Interpreting the fixed model intercept and fixed slope for time

The fixed intercept, $\boldsymbol{\gamma_{000}}$ = 92.47, is the model predicted baseline fidelity score for providers with average scores for both EBPAS factors, who have no experience serving at-risk families, implementing a structured parenting program, or training providers in structured parenting interventions, and who have no training themselves in structured parenting interventions or evidence-based interventions. The fixed slope for time, $\boldsymbol{\gamma_{100}}$ = .001, indicates a statistically significant growth rate of fidelity over time when adjusting for the other covariates ($p = .0280$). So, we may say that this model *predicts* that a provider with average scores on both EBPAS factors and no experience or training (as outlined above) will display an increase in fidelity of .001 percentage points for each day since their first visit. Of course, actual observed providers will deviate from this predicted value, due to random variation and imperfection in the fit of the model.

Interpreting the cross level direct and interaction effects

The fixed beta coefficient for the EBPAS positivity factor, $\boldsymbol{\beta_{2jk}}$ = 0.73, indicates that a one point increase on that factor corresponds on average to fidelity scores that are 0.73 percentage points higher, controlling for covariates ($p = .0019$), while the corresponding fixed beta coefficient for the EBPAS negativity factor, $\boldsymbol{\beta_{3jk}}$ = -0.51, indicates that a one point increase on that factor corresponds on average to fidelity scores that are 0.51 percentage points lower, controlling for covariates ($p = .0121$). Experience serving at risk families was associated with fidelity scores when controlling for other predictors. Specifically, the coefficient, $\boldsymbol{\beta_{6jk}}$ = 1.29, indicates that having more than one year of experience serving at risk families was associated, on average, with fidelity scores that were 1.29 percentage points higher (p = .0160), and $\boldsymbol{\beta_{7jk}}$ = 1.30, indicates that having some, but less than one year, of experience serving at risk families was associated, on average, with fidelity scores that were 1.30 percentage points higher than providers with no work experience (p = .0338).

Providers who had more than one year of experience training others in structured parenting interventions had fidelity scores that were, on average, 1.22 percentage points lower than those with no such experience (p = .0031), while having some, but less than one year of experience was not associated with fidelity. Having some, but less than one year of experience implementing structured parenting programs was associated with greater fidelity growth compared with none, ($\boldsymbol{\beta_{13jk}}$ = .002, p = .0243). Having learned a structured parenting program in the past was also associated with fidelity growth rates when controlling for covariates. The coefficient for the interaction term of this variable with time, $\boldsymbol{\beta_{15jk}}$ = -.003, indicates that providers who had learned structured parenting had fidelity growth rates that were, on average, .003 percentage points smaller than those who had not (p = .0008). This may suggest a ceiling effect, given that overall fidelity is already high. Having prior training in evidence-based interventions was not associated with baseline fidelity when controlling for other predictors, but it was associated with fidelity growth. The coefficient for the interaction term of this variable with time, $\boldsymbol{\beta_{17jk}}$ = .003, indicates that providers with prior training in EBI have fidelity growth rates that are, on average, 0.003% greater than those without prior training (p = .0011).

The beta coefficients reported above have some utility in assessing the predictive power of the variables in the model, but the fact that they are scale specific has both advantages and disadvantages (Aguinis et al., 2013). Scale specificity aids in the direct interpretation of the parameter estimate's predictive relationship with the outcome of interest, but precludes direct comparisons of predictive power between predictors on different scales. In the case of these analyses, this means, for example, that parameter estimates for the two EBPAS factors (positivity and negativity) would be directly comparable. However, the parameter estimate for an EBPAS factor could not be compared directly in this way with the parameter estimate for a non-EBPAS variable. Of the two EBPAS factors, positivity has a greater degree of predictive power, given that its beta coefficient has a larger absolute value ($\boldsymbol{\beta}_{2jk}$ =.73 vs. $\boldsymbol{\beta}_{3jk}$ = -.51). Since both of the variables with interaction terms in the final model are on a yes/no scale, their predictive power can be compared using their respective beta coefficients. In doing so, we can see that, while the two variables have opposite relationships with fidelity growth, their predictive power is similar ($\boldsymbol{\beta}_{9jk}$ = -.003 $\boldsymbol{\beta}_{11jk}$ = .003).

Interpreting the random component

The inclusion of covariates did not result in any major changes to the variance component estimates. Both the provider level and organization level intercept estimates decreased slightly with the inclusion of covariates and interaction terms, indicating a slight improvement in model fit and slightly greater proportion of *explainable* variance accounted for. As Singer (2002) notes, proportion of explainable variance accounted for is not the same as an R^2 statistic, however, which is the proportion of *total* variance explained: "If the amount of variation between individuals is small, we might be explaining a large amount of very little!" (p. 162). In this case, the inclusion of covariates seems to have explained a small proportion of a small amount of variability, though both organization level intercept variance and provider level intercept variance were still statistically significant (Org level: $\boldsymbol{\varphi}^2_0$ = 4.54, p < .0001; Provider level: $\boldsymbol{\tau}^2_0$ = 10.78, p < .0001). Organizational level intercept variance decreased from 4.87 in Model 2 to 4.54

in Model 5, while provider level intercept variance decreased from 11.71 to 10.78. Organization level slope variance was once again nonsignificant ($p = .4934$), while provider level slope variance was still statistically significant, but arguably practically negligible ($\boldsymbol{\tau}^2_1 = .009$, $p < .0001$). Provider level slope intercept covariance was statistically significant, indicating that providers with higher baseline scores had very slightly smaller growth rates on average ($\boldsymbol{\tau}_{02} = -0.00002$, $p < .0001$). It is unsurprising that the residual variance for this model has remained almost the same as it was in the unconditional linear growth model (Model 2 $\boldsymbol{\sigma}^2 = 45.69$, Model 5 $\boldsymbol{\sigma}^2 = 45.50$) because, as Singer (2002) states, "it is difficult for a person-level covariate to help explain *within* person variability" (p. 160).

Interpreting model fit statistics

Each of the three major fit statistics favored the final conditional linear growth model (Model 5, the final multivariable model) over the unconditional linear growth model (Model 2, the model including only time as a predictor). The -2LL statistic decreased from 97810.2 to 96643.3, while the AIC decreased from 97889.2 to 96693.3and BIC from 97917.5 to 96772.0.

Chapter 4: Discussion

Summary

The purpose of this study was to explore SafeCare implementation fidelity trajectories and to investigate factors that may impact those trajectories, with a particular focus on the relationship between EBP attitudes and fidelity. I had two main hypotheses. First, I hypothesized that SafeCare fidelity would display positive growth, but that that growth would be limited by a ceiling effect. This hypothesis was consistent with the analytic findings. Second, I hypothesized that positive EBP attitudes would relate to higher fidelity overall but not to fidelity growth, while negative attitudes would relate to lower fidelity overall and lower fidelity growth. Each component of this latter hypothesis was consistent with analytic findings, except for the relationship between negative attitudes and fidelity growth. No relationship between these two was observed.

Interpretation

This paper describes data collected through the ongoing implementation of SafeCare across 172 partner organizations, and the findings indicate, first and foremost, that SafeCare fidelity is stable and high. SafeCare fidelity begins high at baseline (93.85% on average), and displays a statistically significant trend of positive linear growth. Coaching supports have been shown in systematic review to be supportive of sustained high fidelity, and it is possible that the stable, high fidelity observed here is explained in part by the intensive coaching that is a part of NSTRC's implementation approach (Bartley et al., 2017). However, SafeCare fidelity supports are more intensive early in a provider's implementation of the intervention than they are later on. If intensive coaching supports were necessary to *maintain* high fidelity, then it may reasonably be expected that fidelity would begin to decrease as the intensity of coaching support tapered off.

This does not appear to be the case, however. These findings, therefore, are more similar to those of Clements et al. (2015), who found that fidelity remained stable and high after the total cessation of fidelity support, than those of Chiapa et al. (2015), who took declining fidelity as an indication of the importance of continuous, long term fidelity support.

The major predictors of interest, EBPAS positivity and EBPAS negativity, were each found to relate to fidelity overall in the manner hypothesized, with positive attitudes being associated with higher fidelity and negative attitudes associated with lower fidelity. With regard to baseline fidelity, these findings were consistent with the conditional linear growth models including time by predictor interaction terms, where positive attitudes were associated with higher baseline fidelity and negative attitudes associated with lower baseline fidelity. Again, consistent with the research hypotheses, positive EBP attitudes were *not* associated with increased fidelity growth, possibly due to a ceiling effect. Inconsistent with the research hypotheses, however, negative EBP attitudes were not associated with any differences in fidelity growth. On the whole, these findings provide some support for the idea that provider attitudes toward EBP are related to implementation fidelity, but not fidelity growth, though the effect sizes are small.

Race and gender were not related to either mean fidelity scores or change in fidelity over time. Having a graduate education was associated with higher baseline fidelity and mean fidelity when compared with only having a high school education, though no relationship between provider education and fidelity growth was observed. There were no observable differences in mean fidelity, baseline fidelity, or fidelity growth based on the number of years a provider had been on the job, or based on full- versus part-time work status. Greater experience serving at risk families was associated with higher average fidelity, though greater experience serving substantiated families was not. Having at least one year of experience serving at risk families was also associated with higher baseline fidelity. Neither experience training providers to work with high risk families nor experience providing structured parenting interventions were associated with fidelity on average, though having some but less than one year of experience providing structured parenting interventions was associated with higher fidelity growth.

Somewhat surprisingly, having greater than one year of experience training providers in structured parenting interventions was associated with *lower* fidelity on average and at baseline. Neither having received training in structured parenting interventions nor training in other evidence-based interventions was associated with fidelity on average. Having learned a structured parenting intervention was associated with higher baseline fidelity and lower growth, while having learned an evidence-based intervention in the past was associated with higher growth (but not associated with baseline fidelity). With the exception of provider education, each of the variables that was statistically significantly associated with fidelity in prior models (either on average, at baseline, or over time) was still significantly associated with fidelity in the final multivariable model, indicating that the relationships observed in prior models were not eliminated by controlling for other significant predictors.

How do these findings comport with the existing literature?

The overall positive trend in fidelity growth is consistent with the findings of Chaffin et al. (2016), which is the study that most closely mirrors that which was undertaken here. They are also consistent with the findings of at least two other studies where positive growth in fidelity was observed over time (Forgatch & DeGarmo, 2011; Schaper et al., 2016). The general lack of evidence found here for a relationship between demographic predictors such as race and sex is consistent with the findings of several previous studies (i.e. Whitaker et al., 2012; Bearman et al., 2013). Bartley et al. (2017) note that results in the literature are mixed with regard to the relationship between time on the job and fidelity. The lack of observed relationship here is not consistent with the findings of Beidas et al (2015) who observed an inverse relationship between work experience and fidelity, nor is it consistent with those of Taylor et al (2015), who found greater work experience was associated with higher fidelity. However, these findings are complicated by the fact that, although there was no observed relationship between years on the job and fidelity in these analyses, there were several variables pertaining to more specific aspects of work experience that were related to fidelity. To add yet more complexity, service experience with at risk families was positively associated with fidelity, while experience training providers in structured parenting programs was negatively associated. With regard to EBP attitudes and fidelity, though the effect sizes are small, these findings are consistent with those of Sijercic et al (2020), who found in the context of a randomized trial that positive EBP attitudes are correlated with higher fidelity.

Limitations

While the data analyzed here represent a unique opportunity to explore both EBP implementation fidelity trajectories and their predictors at a large scale in a child welfare context, there are some inherent limitations to these data as well. For example, while the nesting structure, as analyzed here, is sessions within providers within organizations, as mentioned above, there is actually another possible level of nesting that could potentially contribute to some of the variability in fidelity scores: namely, coaches. Individual SafeCare sessions are conducted by providers and then scored by SafeCare coaches, who are either housed at NSTRC or within the implementing partner organization. Given that coaches act as raters in this capacity, it is possible that differences between coaches in terms of scoring approach may also influence fidelity scores to some extent. For example, some coaches may be more lenient than others, or some coaches may watch more carefully for certain fidelity components than others. It is beyond the scope of this project to explore issues related to inter-rater reliability of the SafeCare fidelity rating scale, but future efforts to do so could potentially contribute to a more precise understanding of the variability in SafeCare fidelity. One further limitation pertains to the available data at the provider and organization levels. While the EBPAS instrument provides useful insight into provider attitudes, the study could be improved through the inclusion of more robust measures of other provider characteristics, as well as characteristics at the organizational level. The inclusion of a validated measure of organizational climate, for example, could potentially explain more variation in fidelity, and also facilitate moderational analyses between individual attitudes and organizational climate as they relate to fidelity.

A third limitation is related to the measure of fidelity itself. As noted above, there are various ways to conceptualize and measure fidelity. For the SafeCare model, fidelity is scored as a checklist of behaviors that should occur during a session, and the percentage of behaviors that actually occurred is the index of fidelity for that session. Within that measure, all behaviors listed are given the same weight. However, there are some components of the measure that are likely more important than others. Some items cover general process components (i.e. “exchanges appropriate initial greeting”), while others cover critical content components (i.e. “explains skills/behaviors to parent”). While both of these items are given equal weight in calculating the final fidelity score, the latter constitutes a more central component of the SafeCare model. In other words, it may be understood as one of the “active ingredients” that render SafeCare effective in improving parenting skill. Further, while the measure assesses provider adherence to the model, it does not fully assess what Proctor et al. (2011) term “quality of delivery,” or what

Webb et al. (2010) term "competence." Though there are slight differences in the use of these terms, they generally refer to the level of skill with which the provider delivers the intervention. Both Feely et al. (2018) and Bond and Drake (2020) advocate that fidelity measures should focus not only on model adherence, but issues related to quality of delivery (or competence) as well, and the measure under study here could be significantly improved by including a more robust treatment of these factors.

Practical significance

While the final multivariable linear growth model contains several statistically significant predictors of baseline fidelity and fidelity trajectories, these findings must be interpreted in light of the small effect sizes observed. A positive one unit difference in EBPAS positivity was associated with a positive difference in baseline fidelity of 0.73 percentage points, and a positive one unit difference in EBPAS negativity was associated with a negative difference in fidelity of 0.51 percentage points. Given that the EBPAS scale ranges only from 1 to 5, and 14 out of 15 items on the scale have a standard deviation in this sample of less than 1, it would be reasonable to argue that, despite the statistical significance of these results, the clinical significance of the effect of provider attitudes on fidelity is minimal. For each of the other predictors in the multivariable model, the situation is similar, with parameter estimates for 2- or 3-level categorical variables ranging between $|0.19|$ and $|1.30|$ percentage points. In fact, even time arguably does not have a practically significant relationship with fidelity, with small effect sizes in all models where it is included and an indication (calculated above on p. 62) that only 5% of the variability in fidelity within providers is accounted for by the inclusion of time in the model. Ultimately, in keeping with the suggestion of Kim et al. (2018) that the difference in outcomes related to extremely high fidelity as opposed to merely sufficient fidelity may be negligible, it might be reasonable to argue here that there are no *practically* significant predictors of fidelity in these data at all.

Some of the most practically significant observations can be found before the inclusion of predictors altogether, in the random variance components of the unconditional means model. They show that while 15% of the variability in fidelity is attributable to clustering at the provider level, and 8% of the variability in fidelity is attributable to clustering at the organization level, 72% of the variability is not explained by clustering of either kind. This means that *within* provider and *within* organization variability is much greater than that *between* providers and *between* organizations. Further, given that the variability in fidelity overall is somewhat limited, the provider level explanatory variables in this model can, at best, explain only a portion of the minimal variability that is occurring in the first place. Caution is therefore warranted in interpreting the findings regarding specific predictors in this context. From a research standpoint, a data source containing highly variable fidelity which is strongly related to differences in provider and organizational characteristics could be seen as desirable if it yielded scientific insights. However, from the standpoint of program evaluation, these findings are actually quite encouraging. Implementing organizations like NSTRC undertake significant efforts to systematize program implementation and coaching support *for the precise purpose of reducing the effect that differences between organizations and providers may have on implementation fidelity and other implementation outcomes.* These findings might suggest, therefore, that regarding fidelity, *those efforts are having the desired effect.* Or, stated differently, these data might suggest that it is the initial training methods implemented that drive the fidelity outcomes observed here, rather than individual or organizational differences.

In order to illustrate this point, it may be helpful to imagine that the findings were radically different: that fidelity was highly variable, with providers frequently not meeting the passing threshold on the fidelity rating scale. If SafeCare was not being implemented with fidelity, the resources being committed to achieving and maintaining fidelity (training, coaching, and monitoring) would be wasted. Worse yet, it would be impossible to evaluate the impact of SafeCare in any meaningful way, given that, without reasonably high fidelity*, we could not say that SafeCare had actually been implemented at all.* A situation such as this would not only preclude meaningful evaluation, it would make it difficult to rule out the possibility that ineffective or even harmful programming is being implemented with clients. So, results such as these (which may seem somewhat unexciting) should actually be understood as an indication of implementation success, at least with regard to implementation fidelity. Implementation fidelity, however, is not the only outcome that matters.

Placing these findings in context: is fidelity enough?

Implementation fidelity can be seen as a necessary, but not sufficient, condition for successful program implementation. Throughout the development of the field of implementation science, fidelity has been conceptualized in various ways. In some instances (as is the case here) fidelity is understood as adherence to the prescribed model parameters as identified within an inventory or check list. In other cases, other factors such as quality of delivery, dosage, or even client responsiveness are included under the umbrella of fidelity. In addition to the significant diversity in what implementation scientists and program evaluators mean when using the term fidelity, there are also a broad range of measurement approaches through which those diverse concepts are operationalized (Mowbray et al., 2003; Berkel et al., 2011; Bartley et al., 2017). Whatever the measurement regime, and whether or not other factors such as quality, dosage, and client responsiveness are understood as components of fidelity, all of these outcomes represent important components of (or necessary conditions for) successful implementation (Proctor et al., 2011). Further, as Rohrbach et al. (2010) argue, it is likely that different dimensions of implementation each relate to one another and function together to produce outcomes. It is possible that fidelity could be high in the observed sample, but that other implementation components could exhibit considerably less favorable conditions. In these cases, implementation success and positive end-user outcomes are unlikely to be achieved, despite nominally high and stable program fidelity.

For example, Whitaker et al. (2012) found in a sample of 295 service providers across 50 agencies who were trained in SafeCare that trainees scored highly on performance measures, including quizzes, roleplays, and in-field fidelity monitoring. The *quantity* of implementation, however, was extremely low, with approximately 25% of trained providers actually implementing the SafeCare intervention in real-world practice following their training. The authors hypothesize that this high quality, low quantity outcome with regard to SafeCare implementation could be explained by system-level barriers to implementation (what Aarons et al. would call in their 2011 article *outer context* factors that influence implementation success or failure). While providers were enthusiastic about the intervention, and organizational buy-in was present (as evidenced by the fact that nearly 300 providers were trained), challenges related to the size of the state in which implementation was occurring, complexities in the referral process, and insufficient knowledge of SafeCare among key external stakeholders contributed to the lack of implementation *quantity*. No matter how high the fidelity is in the observed sample, if providers do not implement the intervention widely, then many potential end-users will not reap the benefits of an EBP. This example focuses on the ways in which high fidelity could co-occur with a lack of full implementation success due to low *dosage,* but there are myriad other implementation components that constitute necessary preconditions for implementation success. In the absence of these conditions, even high fidelity implementation may fail to yield positive end-user outcomes.

Fidelity measures may therefore be more meaningfully understood when interpreted in tandem with other implementation outcomes such as dosage, quality of delivery, and client responsiveness (to the extent that these are not covered within the fidelity measure itself), as well as acceptability, appropriateness, costs, penetration, and sustainability (Proctor et al., 2011). Additionally, outcomes defined by Proctor et al. (2011) as service outcomes, such as efficiency, safety, effectiveness, equity, patient-centeredness, and timeliness each likely interact with implementation fidelity to impact overall implementation success. Further, the ultimate purpose of ensuring fidelity (and, for that matter, implementing EBP's in the first place) is to maximize the likelihood of positive program impact for end-users, such as client satisfaction and performance on target outcomes (i.e. here, parenting behaviors and skills in the short term, and child maltreatment prevention in the long term). The importance of fidelity, therefore, would certainly be better understood if these findings were interpreted in the context of service outcomes and end-user outcomes as well.

Directions for future research

The causal mechanisms that contribute to high or low fidelity, much like the mechanisms that contribute to implementation success or failure overall, are complex and context dependent. Differences in implementation setting, in the specific requirements of the EBP implemented, and in the characteristics of the individuals implementing the intervention all interact dynamically to produce implementation outcomes, including fidelity. These implementation outcomes, in turn, exert effects on end-user outcomes through a complex set of direct, mediating, and moderating relationships with service outcomes like equity and timeliness. According to Berkel et al. (2011), "Studying implementation variables in isolation limits our understanding of how they influence each other and their relative influence on outcomes" (p. 28). In their article, "Putting the pieces together: an integrated model of program implementation," these authors propose a theoretical framework for understanding the ways in which implementation components interact to produce outcomes (Berkel et al., 2011). Noting the lack of distinction between concepts like program fidelity, program adaptation, and quality of delivery, they exhort implementation scientists to develop a greater degree of clarity and consensus regarding the definition and operationalization of key implementation components. While there is a growing body of evidence that implementation components impact outcomes in various ways, the complex causal mechanisms through which these effects occur are not yet fully understood. With greater clarity in the definition of key terms, it would be possible to explore more precisely formulated questions about the ways in which these key components interrelate on the causal pathway leading to end-user outcomes.

While efforts to systematize the definition and measurement of implementation components and outcomes are an important component of implementation science, it is to be expected that the complex and dynamic challenges of clinical practice in community settings will make this difficult. Still, this is a necessary task if the ultimate mission of widely disseminating rigorously tested, effective programs to improve health and welfare in vulnerable populations is to be achieved. In order to do so in the context of any given program implementation, it will be critical to understand fidelity within the broader context of other key implementation components, the capacities and priorities of the implementing organization, and the population the implemented program serves. Therefore, the importance of SafeCare implementation fidelity could be better understood in the context of other implementation outcomes, such as dosage and quality, as well service outcomes and end-user outcomes. In the future, research endeavors that simultaneously track these other factors could yield data that can provide more interpretive depth and better inform service delivery.

References

Aarons, G. A. (2004). Mental health provider attitudes toward adoption of evidence-based practice: The Evidence-Based Practice Attitude Scale (EBPAS). *Mental Health Services Research, 6*(2), 61-74. doi:10.1023/B:MHSR.0000024351.12294.65

Aarons, G. A., Cafri, G., Lugo, L., & Sawitzky, A. (2012). Expanding the Domains of Attitudes Towards Evidence-Based Practice: The Evidence Based Practice Attitude Scale-50. *Administration and Policy in Mental Health and Mental Health Services Research, 39*(5), 331-340. doi:10.1007/s10488-010-0302-3

Aarons, G. A., Glisson, C., Hoagwood, K., Kelleher, K., Landsverk, J., & Cafri, G. (2010). Psychometric properties and U.S. National norms of the Evidence-Based Practice Attitude Scale (EBPAS) [American Psychological Association doi:10.1037/a0019188]. Retrieved

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a Conceptual Model of Evidence-Based Practice Implementation in Public Service Sectors. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 4-23. doi:10.1007/s10488-010-0327-7

Aarons, G. A., & Sawitzky, A. C. (2006). Organizational culture and climate and mental health provider attitudes toward evidence-based practice. *Psychological Services, 3*(1), 61-72. doi:10.1037/1541-1559.3.1.61

Aarons, G. A., Sommerfeld, D. H., Hecht, D. B., Silovsky, J. F., & Chaffin, M. J. (2009). The impact of evidence-based practice implementation and fidelity monitoring on staff turnover: Evidence for a protective effect [American Psychological Association doi:10.1037/a0013223]. Retrieved

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-Practice Recommendations for Estimating Cross-Level Interaction Effects Using Multilevel Modeling. In (Vol. 39, pp. 1490-1528).

Bartley, L. H., Bright, C. L., & DePanfilis, D. (2017). Contributors to Fidelity of Child Welfare-Related Interventions: A Review. *Journal of Public Child Welfare, 11*(4-5), 433-463. doi:10.1080/15548732.2017.1340222

Bauer, M. S., & Kirchner, J. (2019). Implementation science: What is it and why should i care? *Psychiatry Research*. doi:10.1016/j.psychres.2019.04.025

Bearman, S. K., Weisz, J. R., Chorpita, B. F., Hoagwood, K., Ward, A., Ugueto, A. M., . . . The Research Network on Youth Mental, H. (2013). More Practice, Less Preach? The Role of Supervision Processes and Therapist Characteristics in EBP Implementation. *Administration and Policy in Mental Health and Mental Health Services Research, 40*(6), 518-529. doi:10.1007/s10488-013-0485-5

Beidas, R. S., Marcus, S., Aarons, G. A., Hoagwood, K. E., Schoenwald, S., Evans, A. C., . . . Mandell, D. S. (2015). Predictors of Community Therapists' Use of Therapy Techniques in a Large Public Mental Health System. *JAMA Pediatrics, 169*(4), 374-382. doi:10.1001/jamapediatrics.2014.3736

Bell, B. A., Ene, M., Smiley, W., & Shoenberger, J. (2013). A multilevel model primer using SAS PROC MIXED. *SAS Global Forum 2013*(Paper 433-2013).

Bell, B. A., Smiley, W., Ene, M., & Blue, G. (2014). An intermediate primer to estimating linear multilevel models using SAS PROC MIXED. *SAS Global Forum 2014*(Paper 1869-2014).

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the Pieces Together: An Integrated Model of Program Implementation. *Prevention Science, 12*(1), 23-33. doi:10.1007/s11121-010-0186-1

Bliese, P. D., & Ployhart, R. E. (2002). Growth Modeling Using Random Coefficient Models: Model Building, Testing, and Illustrations. *Organizational Research Methods, 5*(4), 362-387. doi:10.1177/109442802237116

Boedeker, P. (2017). Hierarchical linear modeling with maximum likelihood, restricted maximum likelihood, and fully Bayesian estimation. *Practical Assessment, Research, and Evaluation: PARE, 22*. doi:https://doi.org/10.7275/5vvy-8613

Bond, G. R., & Drake, R. E. (2020). Assessing the Fidelity of Evidence-Based Practices: History and Current Status of a Standardized Measurement Methodology. *Administration and Policy in Mental Health and Mental Health Services Research, 47*(6), 874-884. doi:10.1007/s10488-019-00991-6

Botvin, G. J., Dusenbury, L., Baker, E., James-Ortiz, S., & Kerner, J. (1989). A skills training approach to smoking prevention among hispanic youth. *Journal of Behavioral Medicine, 12*(3), 279-296. doi:10.1007/BF00844872

Bradshaw, C. P., Pas, E. T., Debnam, K. J., Lindstrom Johnson, S., & Rose, C. A. (2015). A Focus on Implementation of Positive Behavioral Interventions and Supports (PBIS) in High Schools: Associations With Bullying and Other Indicators of School Disorder. *School Psychology Review, 44*(4), 480-498.

Chaffin, M., Hecht, D., Aarons, G., Fettes, D., Hurlburt, M., & Ledesma, K. (2016). EBT fidelity trajectories across training cohorts using the Interagency Collaborative Team strategy.

Administration and Policy in Mental Health and Mental Health Services Research, 43(2), 144-156. doi:10.1007/s10488-015-0627-z

Chiapa, A., Smith, J. D., Kim, H., Dishion, T. J., Shaw, D. S., & Wilson, M. N. (2015). The trajectory of fidelity in a multiyear trial of the family check-up predicts change in child problem behavior. *Journal of Consulting and Clinical Psychology, 83*(5), 1006-1011. doi:10.1037/ccp0000034

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2015). Sustainability of a Scale-Up Intervention in Early Mathematics: A Longitudinal Evaluation of Implementation Fidelity. *Early Education and Development, 26*(3), 427-449. doi:10.1080/10409289.2015.968242

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation Science, 4*(1), 50. doi:10.1186/1748-5908-4-50

Durlak, J. A., & DuPre, E. P. (2008). Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation. *American Journal of Community Psychology, 41*(3), 327. doi: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. doi:10.1093/her/18.2.237

Eccles, M. P., & Mittman, B. S. (2006). Welcome to Implementation Science. *Implementation Science, 1*(1), 1. doi:10.1186/1748-5908-1-1

Feely, M., Seay, K. D., Lanier, P., Auslander, W., & Kohl, P. L. (2018). Measuring Fidelity in Research Studies: A Field Guide to Developing a Comprehensive Fidelity Measurement System. *Child and Adolescent Social Work Journal, 35*(2), 139-152. doi:10.1007/s10560-017-0512-6

Fixsen, D. L. S. F. N. K. A. B. R. M. F. F. W. (2005). *Implementation Research: A Review of the Literature*. University of South Florida,.Tampa, Florida: National Implementation Research Network.

Forgatch, M. S., & DeGarmo, D. S. (2011). Sustaining Fidelity Following the Nationwide PMTO™ Implementation in Norway. *Prevention Science, 12*(3), 235-246. doi:10.1007/s11121-011-0225-6

Hoekstra, F., van Offenbeek, M. A. G., Dekker, R., Hettinga, F. J., Hoekstra, T., van der Woude, L. H. V., . . . ReSpAct, g. (2017). Implementation fidelity trajectories of a health promotion program in multidisciplinary settings: managing tensions in rehabilitation care. *Implementation Science, 12*(1), 143. doi:10.1186/s13012-017-0667-8

Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal Impact of Two First-Grade Preventive Interventions on the Early Risk Behaviors for Later Substance Abuse, Depression, and Antisocial Behavior. *American Journal of Community Psychology, 27*(5), 599-641. doi:https://doi.org/10.1023/A:1022137920532

James, A. G., Noltemeyer, A., Ritchie, R., Palmer, K., & University, M. (2019). Longitudinal disciplinary and achievement outcomes associated with school-wide PBIS implementation level. *Psychology in the Schools, 56*(9), 1512-1521. doi:https://doi.org/10.1002/pits.22282

Kim, J., McIntosh, K., Mercer, S. H., & Nese, R. N. T. (2018). Longitudinal Associations Between SWPBIS Fidelity of Implementation and Behavior and Academic Outcomes. *Behavioral Disorders, 43*(3), 357-369. doi:10.1177/0198742917747589

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: an enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science, 10*(1), 155. doi:10.1186/s13012-015-0342-x

McIntosh, K., Mercer, S. H., Nese, R. N. T., & Ghemraoui, A. (2016). *Identifying and Predicting Distinct Patterns of Implementation in a School-Wide Behavior Support Framework.* Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=ED577292&site=eds-live&scope=site&custid=gsu1

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity Criteria: Development, Measurement, and Validation. *American Journal of Evaluation, 24*(3), 315-340. doi:10.1177/109821400302400303

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive–behavioural therapy. *Clinical Psychology Review, 33*(3), 484-499. doi:https://doi.org/10.1016/j.cpr.2013.01.010

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., . . . Hensley, M. (2011). Outcomes for Implementation Research: Conceptual Distinctions, Measurement Challenges, and Research Agenda. In (Vol. 38, pp. 65-76).

Reding, M. E. J., Chorpita, B. F., Lau, A. S., & Innes-Gomberg, D. (2014). Providers' Attitudes Toward Evidence-Based Practices: Is it Just About Providers, or Do Practices Matter, Too? *Administration and Policy in Mental Health and Mental Health Services Research, 41*(6), 767-776. doi:10.1007/s10488-013-0525-1

Ringwalt, C. L., Pankratz, M. M., Jackson-Newsom, J., Gottfredson, N. C., Hansen, W. B., Giles, S. M., & Dusenbury, L. (2010). Three-Year Trajectory of Teachers' Fidelity to a Drug Prevention Curriculum. *Prevention Science, 11*(1), 67-76. doi:10.1007/s11121-009-0150-0

Rohrbach, L. A., Gunning, M., Sun, P., & Sussman, S. (2010). The Project Towards No Drug Abuse (TND) Dissemination Trial: Implementation Fidelity and Immediate Outcomes. *Prevention Science, 11*(1), 77-88. doi:10.1007/s11121-009-0151-z

Rye, M., Torres, E. M., Friborg, O., Skre, I., & Aarons, G. A. (2017). The Evidence-based Practice Attitude Scale-36 (EBPAS-36): a brief and pragmatic measure of attitudes to evidence-based practice validated in US and Norwegian samples. *Implementation Science, 12*(1), 44. doi:10.1186/s13012-017-0573-0

Schaper, A., McIntosh, K., & Hoselton, R. (2016). Within-year fidelity growth of SWPBIS during installation and initial implementation. *School Psychology Quarterly, 31*(3), 358-368. doi:10.1037/spq0000125

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: monitoring treatment fidelity. *Family Process, 39*(1), 83-83. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=rzh&AN=106091388&site=eds-live&scope=site&custid=gsu1

Schoenwald, S. K., Sheidow, A. J., & Letourneau, E. J. (2004). Toward Effective Quality Assurance in Evidence-Based Practice: Links Between Expert Consultation, Therapist

Fidelity, and Child Outcomes. *Journal of Clinical Child & Adolescent Psychology, 33*(1), 94-104. doi:10.1207/S15374424JCCP3301_10

Sijercic, I., Lane, J. E. M., Gutner, C. A., Monson, C. M., & Stirman, S. W. (2020). The Association Between Clinician and Perceived Organizational Factors with Early Fidelity to Cognitive Processing Therapy for Posttraumatic Stress Disorder in a Randomized Controlled Implementation Trial. *Administration and Policy in Mental Health and Mental Health Services Research, 47*(1), 8-18. doi:10.1007/s10488-019-00966-7

Singer, J. (2002). Fitting individual growth models using SAS PROC MIXED. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling Intraindiviual Variability with Repeadted Measures Data: Methods and Applications*. New York: Routledge.

Smith, J. D., Dishion, T. J., Shaw, D. S., & Wilson, M. N. (2013). Indirect effects of fidelity to the family check-up on changes in parenting and early childhood problem behaviors. *Journal of Consulting and Clinical Psychology, 81*(6), 962-974. doi:10.1037/a0033950

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd Edition ed.). Washington, DC: Sage.

Suzuki, S., & Sheu, C. (1999). *Using PROC MIXED in hierarchical linear models: examples from two- and thre-level school-effect analysis, and meta-analysis research.* Paper presented at the Midwest SAS Users Group Conference Proceedings.

Tasca, G. A., Illing, V., Joyce, A. S., & Ogrodniczuk, J. S. (2009). Three-level multilevel growth models for nested change data: A guide for group treatment researchers. *Psychotherapy Research, 19*(4-5), 453-461. doi:10.1080/10503300902933188

Taylor, W. D., Asgary-Eden, V., Lee, C. M., & LaRoche, K. J. (2015). Service Providers' Adherence to an Evidence-Based Parenting Program: What are They Missing and Why? *Journal of Child and Family Studies, 24*(1), 50-56. doi:10.1007/s10826-013-9812-4

Verschuur, R., Huskens, B., Korzilius, H., Bakker, L., Snijder, M., & Didden, R. (2019). Pivotal response treatment: A study into the relationship between therapist characteristics and fidelity of implementation. *Autism, 24*(2), 499-514. doi:10.1177/1362361319876213

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Thereapist adherence/competence and treatment outcome: a meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2), 200-211.

Webster-Stratton, C. H., Reid, M. J., & Marsenich, L. (2014). Improving Therapist Fidelity During Implementation of Evidence-Based Practices: Incredible Years Program. *Psychiatric Services, 65*(6), 789-795. doi:10.1176/appi.ps.201200177

Whitaker, D. J., Ryan, K. A., Wild, R. C., Self-Brown, S., Lutzker, J. R., Shanley, J. R., . . . Hodges, A. E. (2011). Initial Implementation Indicators From a Statewide Rollout of SafeCare Within a Child Welfare System. *Child Maltreatment, 17*(1), 96-101. doi:10.1177/1077559511430722

Wiltsey Stirman, S., A Gutner, C., Crits-Christoph, P., Edmunds, J., Evans, A. C., & Beidas, R. S. (2015). Relationships between clinician-level attributes and fidelity-consistent and fidelity-inconsistent modifications to an evidence-based psychotherapy. *Implementation Science, 10*(1), 115. doi:10.1186/s13012-015-0308-z

Zvoch, K. (2009). Treatment Fidelity in Multisite Evaluation A Multilevel Longitudinal Examination of Provider Adherence Status and Change. In (Vol. 30, pp. 44-61).

Appendix 1: Fidelity Checklists

Provider Fidelity Checklist: Baseline Assessment

Provider ______________________ Session Date ___________________ Family _______________________

Coach ________________________ Module, Session # ________________ Mode (circle) In-Person Audio Video

#	Item	Score			Comments
	Has materials ready				
1	Provider assessment document(s)	+	–	**n/a**	
2	Parent materials	+	–	**n/a**	
3	Materials & other supplies	+	–	**n/a**	
	Opens the session				
4	Exchanges appropriate initial greeting	+	–	**n/a**	
5	Gives module overview	+	–	**n/a**	
6	Gives session overview	+	–	**n/a**	
	Demonstrates appropriate demeanor				
7	Sits facing client	+	–	**n/a**	
8	Communicates empathy, warmth, understanding	+	–	**n/a**	
9	Maintains open posture	+	–	**n/a**	
10	Has good eye contact	+	–	**n/a**	
	Uses active listening techniques				
11	Uses words/expressions to encourage parent to talk	+	–	**n/a**	
12	Uses open-ended questions	+	–	**n/a**	
13	Uses reflective or summarizing statements	+	–	**n/a**	
	Conducts *formal* assessment				
14	Explains purpose (“why”) of assessment	+	–	**n/a**	
15	Explains process (“how”) of assessment	+	–	**n/a**	
16	Assesses required number of activities/rooms/scenarios	+	–	**n/a**	
17	Assesses required variety of activities/rooms/scenarios	+	–	**n/a**	
18	Completes necessary assessment form(s)	+	–	**n/a**	
19	Provides general, positive feedback about assessment(s)	+	–	**n/a**	
	Reviews parent materials				
20	Provides appropriate materials to parent	+	–	**n/a**	
21	Reviews parent materials with parent	+	–	**n/a**	
	Addresses issues that arise during session				
22	Encourages parent to ask questions/express concerns	+	–	**n/a**	
23	Responds to parent questions/concerns	+	–	**n/a**	
24	Uses problem solving when appropriate	+	–	**n/a**	
	Follows appropriate end of session sequence				
25	Summarizes session	+	–	**n/a**	
26	Gives general, positive feedback	+	–	**n/a**	
27	Schedules meeting date/time for next session	+	–	**n/a**	
	TOTAL:				Percent correct = _____ %

Provider Fidelity Checklist: Training

Provider ________________________ Session Date ____________________ Family ________________________

Coach __________________________ Module, Session # __________________ Mode (circle) In-Person Audio Video

#	Item	Score			Comments
	Has materials ready				
1	Provider assessment document(s)	+	–	**n/a**	
2	Parent training materials	+	–	**n/a**	
3	Materials & other supplies	+	–	**n/a**	
	Opens the session				
4	Exchanges appropriate initial greeting	+	–	**n/a**	
5	Gives session overview	+	–	**n/a**	
6	Discusses parent's practice since last session	+	–	**n/a**	
	Demonstrates appropriate demeanor				
7	Sits facing client	+	–	**n/a**	
8	Communicates empathy, warmth, understanding	+	–	**n/a**	
9	Maintains open posture	+	–	**n/a**	
10	Has good eye contact	+	–	**n/a**	
	Uses active listening techniques				
11	Uses words/expressions to encourage parent to talk	+	–	**n/a**	
12	Uses open-ended questions	+	–	**n/a**	
13	Uses reflective or summarizing statements	+	–	**n/a**	
	Conducts *formal* assessments				
14	Explains purpose and/or process of assessment	+	–	**n/a**	
15	Assesses appropriate room/scenario/activity	+	–	**n/a**	
16	Completes necessary assessment form(s)	+	–	**n/a**	
	Trains parent in skills/behaviors				
17	Uses appropriate materials to train parent	+	–	**n/a**	
18	Explains skills/behaviors to parent	+	–	**n/a**	
19	Physically models skills/behaviors	+	–	**n/a**	
20	Has parent practice skills/behaviors	+	–	**n/a**	
21	Uses assessment form to document parent practice	+	–	**n/a**	
22	Provides specific, positive feedback	+	–	**n/a**	
23	Provides specific, corrective feedback	+	–	**n/a**	
24	Repeats SafeCare 4 process until mastery/success or session time expires	+	–	**n/a**	
25	Plans parent's practice before next session	+	–	**n/a**	
	Addresses issues that arise during session				
26	Encourages parent to ask questions/express concerns	+	–	**n/a**	
27	Responds to parent questions/concerns	+	–	**n/a**	
28	Uses problem solving when appropriate	+	–	**n/a**	
	Follows appropriate end of session sequence				
29	Summarizes session	+	–	**n/a**	
30	Asks for and answers parent's questions	+	–	**n/a**	
31	Gives general, positive feedback	+	–	**n/a**	
32	Schedules meeting date/time for next session	+	–	**n/a**	
	TOTAL:				Percent correct = _____ %

Provider Fidelity Checklist: End-of-Module

Provider ____________________ Session Date ____________________ Family ________________________

Coach ____________________ Module, Session # ____________________ Mode (circle) In-Person Audio Video

#	Item	Score			Comments
	Has materials ready				
1	Provider assessment document(s)	+	–	**n/a**	
2	Parent training materials (if needed)	+	–	**n/a**	
3	Materials & other supplies	+	–	**n/a**	
	Opens the session				
4	Exchanges appropriate initial greeting	+	–	**n/a**	
5	Gives session overview	+	–	**n/a**	
6	Discusses parent's practice since last session	+	–	**n/a**	
	Demonstrates appropriate demeanor				
7	Sits facing client	+	–	**n/a**	
8	Communicates empathy, warmth, understanding	+	–	**n/a**	
9	Maintains open posture	+	–	**n/a**	
10	Has good eye contact	+	–	**n/a**	
	Uses active listening techniques				
11	Uses words/expressions to encourage parent to talk	+	–	**n/a**	
12	Uses open-ended questions	+	–	**n/a**	
13	Uses reflective or summarizing statements	+	–	**n/a**	
	Conducts *formal* assessments and further training as needed				
14	Explains purpose ("why") of assessments	+	–	**n/a**	
15	Explains process ("how") of assessments	+	–	**n/a**	
16	Assesses required number of activities/rooms/scenarios	+	–	**n/a**	
17	Assesses required variety of activities/rooms/scenarios	+	–	**n/a**	
18	Completes necessary assessment form(s)	+	–	**n/a**	
19	Provides specific, positive feedback	+	–	**n/a**	
20	Repeats SafeCare 4 to achieve mastery/success	+	–	**n/a**	
21	Determines mastery/success according to rules	+	–	**n/a**	
22	Respectfully communicates to parent if cannot move to next module	+	–	**n/a**	
	Addresses issues that arise during session				
23	Encourages parent to ask questions/express concerns	+	–	**n/a**	
24	Responds to parent questions/concerns	+	–	**n/a**	
25	Uses problem solving when appropriate	+	–	**n/a**	
	Follows appropriate end of session sequence				
26	Summarizes session	+	–	**n/a**	
27	Asks for and answers parent questions	+	–	**n/a**	
28	Gives general, positive feedback	+	–	**n/a**	
29	Provides overview of next module/session	+	–	**n/a**	
30	Completes necessary form for next module (DAC/ Safety Consent)	+	–	**n/a**	
31	Schedules meeting date/time for next session	+	–	**n/a**	
	TOTAL:				Percent correct = _____ %