Prevention Science
https://doi.org/10.1007/s11121-020-01146-y

Student Knowledge Gain Following the Second Step Child Protection Unit: the Influence of Treatment Integrity

Margaret E. Manges[1] · Amanda B. Nickerson[1]

© Society for Prevention Research 2020

Abstract

Treatment integrity is an important yet understudied component of school-based prevention programming, particularly for sensitive topics such as child sexual abuse prevention (CSA). This study examined student- and teacher-level characteristics, including components of treatment integrity, that contributed to greater knowledge gain among students participating in the Second Step Child Protection Unit (CPU). The study was conducted with 1132 students and 57 teachers from four elementary schools enrolled in a randomized controlled trial of the CPU. Students were administered assessments at pre-test, post-test, 6-month follow-up, and 12-month follow-up. Teachers were observed and rated on Content Integrity (CI; adherence to content), Process Integrity (PI; teacher enthusiasm, encouragement, behavior management), and Dose Received (DR; student behavior and interest) when delivering the lessons. Hierarchical linear growth modeling indicated that students who received the CPU made gains in the knowledge of CSA concepts and skills over a 12-month follow-up period. Girls had significantly greater CSA knowledge than boys immediately after the intervention, with gender remaining significant even when accounting for level-3 variables. Older children had better knowledge scores at post-test, but growth over time results revealed that younger students made greater gains. For students in 2nd through 4th grade, CI was more important for post-test outcomes, while for all students, CI and grade taught were important to post-test scores. Teachers of lower grades had students with a faster growth rate on correct responses to vignettes. Implications for CSA prevention programming and future research are discussed.

Keywords Fidelity · Treatment integrity · Randomized controlled trial · Second step · Childhood sexual abuse prevention · School

Childhood sexual abuse (CSA) is an international public health concern, with as many as 25% of girls and 16% of boys experiencing CSA by age 18 (Finkelhor et al. 2014). In addition to the immediate negative impact of CSA (e.g., self-blame, shame (Katerndalh et al. 2005); increased likelihood for developing PTSD, depression, suicide, sexual promiscuity, sexual perpetration, and lower academic achievement (Paolucci et al. 2001)), long-term effects may include low self-esteem, anxiety, depression, anger, substance abuse, eating disorders, sexual difficulties, self-injurious behavior, and revictimization (Briere and Elliott 2003; Daigneault et al. 2017; Irish et al. 2010). School-based prevention education has the greatest likelihood of reaching the largest number of children (Topping and Barron 2009). Evidence indicates that children benefit from these programs by learning concepts and skills, such as recognizing, refusing, and reporting unsafe situations (Davis and Gidycz 2000; Rispens et al. 1997; Tutty 1997; Zwi et al. 2007). Young children are particularly vulnerable to child sexual abuse (Briere and Elliott 2003), and education surrounding these concepts empowers children. However, studies examining CSA prevention programs have methodological limitations, including sampling problems, lack of control groups, failure to assess maintenance of gains, and lack of integrity data (Topping and Barron 2009). Treatment integrity or fidelity of implementation (Gearing et al. 2011) is an issue of particular interest for CSA prevention programs as effective implementation is essential to improving the lives of those receiving the treatment (Fixsen et al. 2005). This study examined the relationship between components of treatment integrity and their impact on outcomes over four waves of data collection for students receiving the Second Step CPU's six lessons in recognizing and refusing unsafe situations (Committee for Children 2014).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11121-020-01146-y) contains supplementary material, which is available to authorized users.

✉ Margaret E. Manges
memanges@buffalo.edu

[1] Alberti Center for Bullying Abuse Prevention, The University at Buffalo, State University of New York, 428 Baldy Hall, Buffalo, NY 14260-1000, USA

Published online: 20 July 2020

Treatment Integrity

Although the development of evidence-based programs (EBPs) has improved over recent years, the study of implementation fidelity has lagged behind (Fixsen et al. 2005). Treatment integrity is the extent to which a program or intervention is implemented as intended (Dane and Schneider 1998). Without assessing treatment integrity, variations in intervention effects due to integrity may be lost completely (Collier-Meek et al. 2018), and inconclusive conclusions may be drawn (Perepletchikova 2011). In Dane and Schneider's (1998) meta-analysis, 39 of 162 (~23%) articles clearly documented integrity procedures and only 13 studies considered variations of treatment integrity in examining prevention program effectiveness. A more recent meta-analysis noted that even high-quality education journals inconsistently report integrity scores and less than 70% of articles reported information on integrity (Swanson et al. 2013). Of the studies that provided integrity data, less than 10% included information regarding the quality of the intervention (Swanson et al. 2013). There is a lack of data on treatment integrity for CSA prevention programs (Topping and Barron 2009). It is especially critical to examine treatment integrity in CSA prevention because teachers may lack the prerequisite knowledge or skills to discuss these issues with their students (Marquez-Flores et al. 2016) and may not view mental health prevention or intervention as part of their duties (Reinke et al. 2011).

Components of Treatment Integrity

Treatment integrity has various conceptualizations and methods of assessment, although it typically includes Content Integrity (CI) and Process Integrity (PI). CI, also referred to as adherence (Hagermoser Sanetti and Fallon 2011; Southerland et al. 2018), is the degree to which an intervention is implemented as intended. PI, or competency (Hagermoser Sanetti and Fallon 2011; Southerland et al. 2018), measures "the extent to which facilitators encouraged student participation, utilized appropriate behavior management strategies demonstrated enthusiasm, and managed time adequately" (Gullan et al. 2009, p. 4). Some conceptualizations also include treatment differentiation or the relative effectiveness of treatment components without other interventions being implemented in addition (Perepletchikova and Kazdin 2005). It is important to assess not only the implementation of the intervention but also the participant response (Dane and Schneider 1998; Gullan et al. 2009). Beyond the typical focus on dose delivered (i.e., implementation of program components), Dose Received (DR) refers to the extent to which the participants are actively engaged in the program (Gullan et al. 2009; Nelson et al. 2012). For example, Gullan et al. (2009) conceptualize and assess DR by student behavior, interest, and enthusiasm as indictors of participant engagement in the process (Gullan et al. 2009).

Higher levels of treatment integrity, particularly DR and PI, are associated with better outcomes (Durlak and DuPre 2008). In a study of treatment integrity in schools examining children with autism spectrum disorder, integrity was maintained only around 50% of the time (Mandell et al. 2013). Only students in low- and high-treatment integrity conditions experienced a significant gain, likely due to the level of experience of the teachers administering the lessons (Mandell et al. 2013). Therefore, it is also important to examine unique teacher-level variables, such as years of teaching experience, that could potentially affect intervention outcomes.

Second Step Child Protection Unit

Through an ecological approach, CSA prevention emphasizes the importance of parent, child, professional, and public education by making an impact on policies, laws, and social norms (Kenny and Wurtele 2012). All staff implementing the lessons complete online training and then administer 6 weeks of comprehensive lessons to students in the classroom (Committee for Children 2014). Results from a randomized controlled trial of the CPU reveal that students in the intervention condition had significantly higher post-test scores on CSA prevention concept knowledge and ability to recognize, report, and refuse unsafe touches than students in the control schools after controlling for baseline scores, with small to medium effect sizes ranging from $\eta^2 = 0.001$ to $\eta^2 = 0.07$ (Nickerson et al. 2019). Child age and gender moderated the findings, with children in younger grades showing greater gains and girls achieving better outcomes than boys (Nickerson et al. 2019). Teachers in the intervention schools also had increased knowledge of CSA, more positive attitudes towards reporting CSA, and improved perceptions of teacher-student relations compared with those in the control group (Kim et al. 2019). Although fidelity of implementation was reported in previous studies (average of 81% observed adherence to specific steps in lesson manuals; Nickerson et al. 2019), it was not examined with respect to outcomes. The CPU contains multiple aspects, including training for teachers and administrators, policy changes, work with parents, and classroom implementation (Committee for Children 2014), although the current study focused primarily on the core features of lesson implementation.

Current Study

This study explored which components of treatment integrity predicted greater knowledge gain among students following the Second Step CPU intervention. Aims were to (a)

determine which student-level characteristics, such as age, gender, school, and grade, influenced greater knowledge gain among intervention schools and (b) determine which teacher-level characteristics, such as CI, PI, DR, grade taught, and years of experience, influenced baseline knowledge and knowledge gain among students in intervention schools. It was hypothesized that (a) higher CI, PI, and DR would lead to higher intercept (post-test) scores at 12-month follow-up, as measured by the *Children's Knowledge of Abuse Questionnaire-Revised* (*CKAQ*; Tutty 1997) and the *What-If Situations Test-III-R* (*WIST*; Wurtele et al. 1988; Wurtele et al. 1989); (b) higher CI, PI, and DR would lead to greater knowledge gain among students at 12-month follow-up as measured by the CKAQ and WIST (Durlak and DuPre 2008); and (c) other teacher variables, such as years of teaching experience, would not account for as much statistical significance in intercept scores or knowledge gain as treatment integrity variables. Outcome variables were selected based on prior research measuring efficacy of treatment integrity through gains in knowledge or skills (Mandell et al. 2013; Nelson et al. 2012).

Method

The data from this study were drawn from a cluster randomized controlled trial (RCT) to assess the effectiveness of the Second Step CPU. Eight elementary schools randomly selected from the district's 11 elementary schools and matched based on grade level (PreK-2, K-5, 3–5), school size, racial/ethnic diversity, and percent of students receiving free and reduced lunch were randomly assigned to the intervention or wait-list control group. The school district was in a suburban area of Western New York serving just over 11,000 students. Approximately 61% of students were White, 14% Black, 15% Hispanic, 7% multiracial, and 3% American Indian, Alaskan, Asian, or Pacific Islander; 45% were economically disadvantaged.

Because this study examined integrity of implementation, only data from the teachers and students in the intervention condition were included. Interviews conducted with school principals prior to the start of the intervention revealed that, per state law, all teachers and school staff had received the mandatory 2-h training in child abuse identification and reporting prior to being certified. This is a one-time state requirement, so for many teachers, the length of time teaching was the length of time since completing the course. In focus groups completed after the first year of the implementation (Allen et al. 2019), teachers indicated that the training included in the curriculum was pivotal in increasing their awareness of CSA and in preparing them not only to teach the curriculum but also to recognize and report CSA. No school had CSA prevention programming implemented by classroom teachers, although in two of the eight schools the counselor used self-created materials to talk about touching safety. Lessons were administered in the classroom by teachers (or co-led by the school counselor), with pre-kindergarten (PK) and kindergarten (K) students receiving short daily lessons and older students (grades 1 through 4) receiving 30- to 45-min lessons once per week. Lessons were delivered in a developmentally appropriate way, utilizing puppets, music videos, stories, and structured discussions.

Participants and Procedure

The student participants of this study included children in PK through grade 4 at baseline. All schools used a waiver of active consent (passive) procedure approved by the University Institutional Review Board. A total of 1132 students and 57 teachers completed all waves of data collection. The ages of students at pre-test ranged from 4 to 12 years old ($M = 7.19$, $SD = 1.58$), with 49% of the study being male. In addition to 130 students whose parents initially opted their children out, 263 students did not complete the follow-up surveys due to absence, no longer in attendance, or not giving assent. Some classrooms were unable to be observed due to student issues and concerns (e.g., special education classroom with disruptive behavior). Although the data collected at the fourth time point included fifth grade teachers, this sample was not used as these teachers did not implement the lessons at pre-test, resulting in the final teacher N of 57. The teachers in this study taught grades PK through 4th grade, with a range of 1 to 30 years of experience, and teachers were aged from 25 to 55 years old.

Staff completed two modules of self-paced online training prior to administering the lessons. One module (75–90 min), intended for all staff, taught participants to recognize, respond to, and report abuse and neglect. The other module (75–90 min) was designed to prepare the teachers how to administer the lessons, overcome any discomfort brought upon by sensitive lesson material, and engage families using provided materials (Committee for Children 2014). Staff were provided with time during the school day (e.g., in lieu of faculty meeting) to complete the modules. The project coordinator verified completion for each teacher via the Second Step Online Administrator Dashboard, through allowed training invitations and reminders to be sent and to track training progress. All teachers completed the two required modules and were compensated with $50. Upon completion of the staff training, teachers implemented the intervention during the school day, with variation among teachers on what time the lessons were implemented (some teachers delivered the lessons in the morning, others chose the afternoon). Student lessons lasted 6 weeks, with 6 lessons in total. Lesson content focused primarily on the development of knowledge and skills related to CSA, utilizing stories, discussion, videos, music, and puppets (for young children). Lessons were administered for two

consecutive years, with the first wave of implementation occurring during the 2017–2018 school year and the second wave occurring during the 2018–2019 school year. Students in the first wave were in PK through 4th grade, and these students were followed to the next year, as K-5th grade students, respectively. Within approximately 1 week following the pre-test assessment, teachers began implementing the lessons within the classrooms. Research staff observed and collected treatment integrity data on one-third (2/6) of the lessons that were taught in the first year of the program, which meets or exceeds the standard used in most intervention studies involving fidelity checks (e.g., Ardoin et al. 2016), and accounted for variability in lesson implementation over time (Mowbray et al. 2003).

Student data were collected at four time points: (a) pre-test, (b) post-test, (c) 6-month follow-up, and (d) 12-month follow-up beginning in September of 2017 and completed by January of 2019. Trained graduate students administered the assessment individually, in small groups, and to the whole class depending on the ages and varying needs of students in the classrooms. Data for students were collected either through paper-and-pencil surveys or online through Survey Monkey, and all assessments were read aloud to all students.

Measures

Treatment Integrity Treatment integrity data were collected using the Integrity Monitoring Checklist (IMC), adapted from Gullan et al. (2009) and Mellard (2010). The format for the checklists was uniform, but the specific items were unique to each lesson to reflect the content and lesson components used. Although there was separate content among grades, there was significant overlap between the content in some grades (i.e., 2nd and 3rd, 4th and 5th), and because of this, the IMCs were created accordingly.

Similar to the Gullan et al. (2009) study, treatment integrity data focused on CI, PI, and DR. CI focused on whether or not the classroom teacher followed the key components of the intervention, namely, the script, prompts, and actions laid out by the CPU manuals at the correct time. Scores on this measure were coded from 0 to 2 (0, *missing or incorrect*; 1, *present but needs improvement*; and 2, *present and correct*; there was also a not applicable item; Gullan et al. 2009; Mellard 2010). PI focused on teacher enthusiasm, time management, encouragement of student behavior, and utilization of behavior management strategies (Gullan et al. 2009). Observers rated teachers using a 5-point Likert scale (1 = *did not exhibit the behavior at all or did extremely poorly* to 5 = *exhibited the behavior throughout the lesson with almost all of the students*). CI and PI were measured in five categories: review, introduction, story and discussion, activity/skill practice, and wrap up (with 4–8 CI items for each of the categories).

DR was measured through assessing student interest and enthusiasm and student on-task behavior. Rather than assess the number of students present during the lesson, we wanted to examine whether or not on-task behavior, interest, or enthusiasm played a role in knowledge gain (Gullan et al. 2009). Student interest and enthusiasm was rated on a 5-point Likert scale (1 = *multiple students showing lack of interest in materials; little to no participation and engagement; students are mostly unresponsive to prompts and do not answer questions as they are asked* to 5 = *students are fully engaged throughout the lesson segment*). Student on-task behavior was also assessed on a 5-point Likert scale (1 = *multiple students showing marked distraction or serious disruptive behavior that causes major disruption to content* to 5 = *students are fully attentive throughout the lesson segment with no redirection needed*).

CI was calculated by adding the scores obtained by teachers and dividing that by the total possible score, from which a percent was calculated. PI and DR were calculated by adding the scores obtained on each outcome at each time point (i.e., adding the scores received on encouragement, enthusiasm, time management, organization, behavior management, student interest and enthusiasm, and student on-task behavior) from the review, introduction, story and discussion, and wrap-up sections. The total scores in each category were added together and divided by the total possible score for that outcome.

Observers were graduate research assistants and a faculty member. Over a 2-month period, training sessions were held in which raters learned the Integrity Monitoring Checklists (IMC) and practiced ratings with live and videotaped demonstrations of teachers administering the lessons. Raters then compared scores to assess and refine inter-rater reliability. For the first round of observations, 22.41% of teachers were observed by two raters, and for the second round of observations, 20.33% of teachers were observed by two raters. For the CI scales that were scored from 0 to 3, observers were considered in accordance if ratings were the same. For 5-point Likert scales, raters were considered in agreement if the rating was within 2 points, consistent with Gullan et al.'s (2009) scoring. For example, if observer 1 scored a component as a 4 while observer 2 rated it as a 5, these two ratings would be considered in agreement. Inter-rater reliability (IRR) was calculated by dividing the number of items with agreement by the total number of possible items and multiplying by 100. Based on the present sample, the IRR for the first observation was 86% (CI = 81.4%, PI = 86.2%, DR = 99.1%). IRR for the second round was 90% (CI = 86.6%, PI = 85.6%, DR = 98.5%). In addition, Kappa was 0.94 at time 1 and 0.92 at time 2. For analysis, each lesson had a primary observer; to analyze lessons with multiple observers, the primary observer's ratings were used.

Children's Knowledge of Abuse Questionnaire-Revised (CKAQ; Tutty 1997) The Inappropriate Touch subscale of the CKAQ (α = .87, test-retest reliability = 0.88; Tutty 1997) was

used to assess CSA concept knowledge in students in 2nd through 5th grade. This subscale contains 24 items that measure general concepts of CSA prevention programs. Example questions from this scale include "It's OK to say 'no' and move away if someone touches you in a way you don't like," "You can trust your feelings about whether a touch is safe or unsafe," and "Sometimes someone in your family might touch you in a way you don't like." Correct responses were scored as 1 point, while incorrect or skipped responses were given 0 points. The outcome measure was analyzed as a sum score. Reliability for the current sample was $\alpha = 0.61$.

What-If Situations Test-III-R (WIST; Wurtele et al. 1988; Wurtele et al. 1989) The WIST was used with all students to assess recognition, refusal, and reporting of unsafe situations. It includes 6 vignettes in which adults ask to touch or look at a child's private body parts. Three of the vignettes are inappropriate, while three are appropriate (e.g., a parent or doctor asking to look at the child's body parts after an injury). The WIST assesses if children would say "yes" or "no" to these requests, what they would say to the person making the request, who the child would tell about the interaction, and if the child thinks the situation is "okay" (Wurtele et al. 1989). Questions were modified to be administered as multiple-choice response options, and children's responses to each question in the inappropriate request vignettes are scored from 0 to 2, with higher scores indicating a higher level of skill, yielding a total skill sum score of 24 points (8 points maximum per inappropriate-request vignettes). The internal consistency for the current sample was $\alpha = 0.74$.

Data Analysis

Hierarchical linear growth modeling (HLGM; Bryk and Raudenbush 1992) was used to examine the effects of treatment integrity on students' knowledge scores over time. This analysis consists of two series of growth models: (1) knowledge as measured by the CKAQ (grades 2–5) and (2) knowledge measured by the WIST (PK to grade 5). HLGM is appropriate for this analysis due to the nested nature of repeated measures within students and students within teachers (Bryk and Raudenbush 1992). The data for the current study ranged in missingness from 17 to 26%. Using the missing data classification system, data were determined not to be missing completely at random (MCAR; $p < 0.001$; Little and Rubin 2002). Using the SPSS Missing Data package, data were multiply imputed, and five imputations were conducted. Level-1 variables in the models included knowledge outcomes and time, demonstrating individual-level growth. The time variable in the model was coded as − 1, 0, 1, and 2 corresponding to the four time periods at which students were assessed. The level-1 (time) equation is represented by:

- Level 1:

$$\mathrm{KNOWLEDGE}_{tij} = \pi_{0ij} + \pi_{1ij}(\mathrm{TIME}) + e_{tij}$$

At level-2, knowledge gain at post-test (intercept) and gains over time is a function of student-level characteristics. Both age and gender were grand mean centered so that the intercept could be interpreted as the mean at post-test adjusted for differences between genders and ages. These were included as predictors of knowledge outcomes to allow for a fair comparison among students regardless of age and gender. Gender was coded (*males = 0 and females = 1*), while average age was consistent with actual student age (i.e., 4-year-old students were coded as 4, age was then averaged across the four time points). Level-2 (student) model is represented by:

Level 2:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(\boldsymbol{GENDER}) + \beta_{02j}(\mathbf{AVG_AGE}) + r_{0j}$$
$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(\boldsymbol{GENDER}) + \beta_{12j}(\mathbf{AVG_AGE}) + r_{1j}$$

Finally, level-3 variables in these models included years of experience, grade taught, CI, PI, and DR. The Level-3 (teacher) model is represented by:

Level 3:

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(\mathrm{GRADE}) + \gamma_{002}(\mathrm{EXPERIENCE}) + \gamma_{003}(\mathrm{CONTENT}) + \gamma_{004}(\mathrm{PROCESS}) + \gamma_{005}(\mathrm{DOSE_AVG}) + u_{00j}$$
$$\beta_{01j} = \gamma_{010}$$
$$\beta_{02j} = \gamma_{020}$$
$$\beta_{10j} = \gamma_{100} + \gamma_{101}(\mathrm{GRADE}) + \gamma_{102}(\mathrm{EXPERIENCE}) + \gamma_{103}(\mathrm{CONTENT}) + \gamma_{104}(\mathrm{PROCESS}) + \gamma_{105}(\mathrm{DOSE_AVG}) + u_{00j}$$
$$\beta_{11j} = \gamma_{110}$$
$$\beta_{12j} = \gamma_{120}$$

The ICC for level 2 was 0.1265 and the level-3 ICC was 0.1003, indicating that 12.65% of the variance occurred between students, while 10.03% of the variance occurred between teachers. The ICC for level 2 was 0.3676 (i.e., 36.76% of the variance occurred between students) and the level-3 ICC was 0.2754 (i.e., 27.54% of the variance occurred between teachers). Typically, an ICC around 0.2 is typical of classroom-based clustering effects; therefore, our ICC's of 0.27 and 0.37 give justification to running a multilevel analysis (Musca et al. 2011).

Results

CKAQ There were 3048 records at level 1, 762 students at level 2, and 42 teachers at level 3. On average, students were 8.05 years old (range 6–12 years). The mean knowledge score at time 1 (pre-test) was 18.74 (out of a total of 24). A fully unconditional model was run to determine if the data were appropriate for HLGM and to see if the average knowledge at post-test and the growth rate varied across individual students (see Table 1). The intercept coefficient (knowledge at post-test) was significant (18.74, $p < 0.001$), indicating that post-test CKAQ scores varied across students. The growth rate was significant (0.96, $p < 0.001$), indicating a knowledge gain of 0.96 points per assessment. There was significant variability between students in the initial level of the CKAQ and a significant change in CKAQ outcomes over time. The random effect variance on the intercept (0.77, $\chi^2 = 156.31$, $p < 0.001$) and the growth rate (0.03, $\chi^2 = 69.49$, $p < 0.005$) were significant, suggesting that data were suitable for HLM.

Both gender and average age were added as predictors in the level-2 analysis. Gender was significant after controlling for age ($\beta_{01} = 0.42$, $p = 0.005$; see Table 1), with females scoring higher than males. Age also had a significant association with knowledge gain in response to the baseline score ($\beta_{02} = 0.35$, $p < 0$ .005); each year of increased age was associated with a 0.35 increase in knowledge score. Age ($\beta_{12} = -0.13$, $p = .005$) had significant relations with knowledge gain, with younger students learning more over time. Gender did not relate to knowledge gain ($\beta_{11} = -0.01$, $p = 0.867$). The random effect variance component on the intercept was significant (0.49, $\chi^2 = 114.25$, $p < 0.001$), suggesting that the average knowledge post-test score varied across students. The random effect variance component on the growth rate was significant (0.02, $\chi^2 = 59.74$, $p < 0.05$), suggesting a clustering effect, justifying the need for a three-level model.

Table 1 Linear model of change in knowledge of sexual abuse concepts over time (unconditional and partially conditional models)

	CKAQ			WIST		
Fixed effect	Coefficient	SE	*t*-ratio	Coefficient	SE	*t*-ratio
Fully unconditional model						
Initial status, β_{00}	18.74****	0.16	118.07	17.12****	0.38	44.79
Growth rate, β_{10}	0.96****	0.05	20.97	0.80****	0.12	6.93
Partially conditional model						
Knowledge, π_{0i}						
Intercept, β_{00}	18.76****	0.14	138.31	17.16****	0.20	87.35
Gender, β_{01}	0.42***	0.15	2.83	0.31	0.17	1.84
Age, β_{02}	0.35***	0.12	3.01	1.18****	0.11	10.67
For time slope, π_{1i}						
Intercept, β_{10}	0.95***	0.04	22.98	0.79****	0.08	10.51
Gender, β_{11}	− 0.01	0.07	− 0.17	− 0.08	0.10	− 0.82
Age, β_{12}	− 0.14***	0.05	− 2.90	− 0.34****	0.05	− 7.22
Random effect	Variance	df	χ^2	Variance	df	χ^2
Fully unconditional model						
Initial status, r_{0i}	2.67****	720	2215.28	3.98****	1075	2383.57
Growth rate, r_{1i}	0.09**	720	785.67	0.41	1075	1092.85
Level-1 error, e_{ti}	4.28			10.90		
Partially conditional model						
Initial status, r_{0i}	2.65****	718	2203.11	4.25****	1073	2465.84
Growth rate, r_{1i}	0.09*	718	783.83	0.42	1073	1093.78
Level-1 error, e_{ti}	4.27			10.91		

CKAQ, Children's Knowledge of Abuse Questionnaire; WIST, What-if Situations Test III-Revised; gender (0, female; 1 = , male). *$p < .05$. **$p < .01$. ***$p < .005$. ****$p < .001$

Table 2 Linear model of change in knowledge over time (fully conditional model)

	CKAQ			WIST		
Fixed effect	Coefficient	SE	*t*-ratio	Coefficient	SE	*t*-ratio
Model for initial status, π_{0i}						
Intercept, γ_{000}	18.69****	0.13	144.43	17.17****	0.15	111.60
Grade, γ_{001}	0.26	0.20	1.29	1.61****	0.21	7.62
Experience, γ_{002}	0.01	0.02	0.70	− 0.00	0.02	− 0.08
CI, γ_{003}	0.03**	0.01	2.55	0.05***	0.01	3.37
PI, γ_{004}	− 0.45	0.25	− 1.78	− 0.41	0.36	− 1.14
DR, γ_{005}	0.22	0.36	0.62	− 0.46	0.49	− 0.36
For gender, β_{01}						
Intercept γ_{010}	0.40**	0.15	2.70	0.24	0.17	1.43
For age, β_{02}						
Intercept γ_{020}	0.29	0.17	1.65	0.01	0.20	0.06
Mean growth rate, π_{1i}						
Intercept, γ_{000}	0.94****	0.05	18.26	0.79****	0.07	11.25
Grade, γ_{101}	0.04	0.10	− 0.34	− 0.25*	0.12	− 2.08
Experience, γ_{102}	− 0.00	0.01	− 0.74	0.00	0.01	0.37
CI, γ_{103}	− 0.00	0.00	0.10	− 0.01	0.01	− 1.89
PI, γ_{104}	− 0.07	0.09	− 0.81	0.06	0.17	0.74
DR, γ_{105}	− 0.02	0.24	− 0.17	0.04	0.23	0.18
For gender, β_{11}						
Intercept γ_{110}	− 0.01	0.07	− 0.16	− 0.07	0.10	− 0.48
For age, β_{12}						
Intercept γ_{120}	− 0.16	0.08	− 1.87	− 0.19	0.12	− 1.55
Random effect	Variance	df	χ^2	Variance	df	χ^2
Initial status, u_{00}	0.28****	36	89.46	0.94****	51	192.98
Growth rate, u_{10}	0.01*	36	55.99	0.14****	51	117.04

CKAQ, Children's Knowledge of Abuse Questionnaire; *WIST*, What-if Situations Test III-Revised; gender (0, female, 1, male); *CI*, Content Integrity; *PI*, Process Integrity; *DR*, Dose Received. *$p < .05$. **$p < .01$. ***$p < .005$. ****$p < .001$

In the final model, grade, experience, CI, PI, and DR were added as teacher-level predictors (see Table 2). Only CI ($\gamma_{003} = 0.03$, $p < 0.05$) had a significant effect on student post-test score at level 3. Gender remained a significant predictor of post-test score at level 2 ($\beta_{01} = 0.40$, $p < 0.01$), indicating that girls scored higher than boys at post-test even after accounting for effects at level 3. Teachers with higher accuracy scores produced students with higher knowledge scores at post-test as measured by the CKAQ. After adding level-3 predictors, the growth rate remained significant ($\gamma_{100} = 0.94$, $p < 0.001$), with only age significantly impacting growth rate ($\beta_{12} = -0.16$, $p < 0.05$). Additional variance was explained by the clustered nature of the level-3 model, with younger students exhibiting a faster rate of change.

WIST There were 4528 records at level 1, 1132 students at level 2, and 57 teachers at level 3. On average, students were 7.16 years old (range 4–12 years). The mean knowledge score at the intercept (pre-test) was 17.12 (range 0–24). A fully unconditional model was run to determine if the WIST data were appropriate for HLGM and to see if the average knowledge at post-test and the growth rate varied across individual students (see Table 1). Both the intercept coefficient ($\beta_{00} = 17.12$, $p < 0.001$) and the slope were significant ($\beta_{10} = 0.80$, $p < 0.001$), indicating a knowledge gain of 0.80 points per assessment. There was significant variability between students in the initial level but no significant variability in WIST outcomes over time. The random effect variance on the intercept (0.79, $\chi^2 = 1290.52$, $p < 0.001$) and slope was significant (0.62, $\chi^2 = 339.04$, $p < 0.001$), suggesting that these data were suitable for HLM.

Both gender and average age were added as level-2 predictors in the level-2 analysis (see Table 1). Gender was not significant after controlling for age ($\beta_{01} = 0.31$, $p > 0.05$). However, age had a significant association with knowledge score at post-test ($\beta_{02} = 1.18$, $p < 0.001$); each year increase in age was associated with a 1.17 increase in knowledge score. Age ($\beta_{12} = -0.34$, $p < 0.001$) had significant relations with

knowledge gain over time, with younger students making gains at a faster rate, although initially starting with significantly lower scores. Gender did not relate to knowledge gain ($\beta_{11} = -0.08$, $p = 0.411$). The random effect variance component on the intercept was significant (1.78, $\chi^2 = 306.900$, $p < 0.001$), suggesting that the average knowledge score at post-test varied across students. The random effect variance component on the slope was also significant (0.18, $\chi^2 = 136.24$, $p < 0.001$), suggesting that the average knowledge gain at post-test varied across students. Overall, variance decreased by adding the level-2 model.

In the final model, grade, experience, CI, PI, and DR were added as teacher-level predictors of student knowledge gain on the WIST (see Table 2). In this model, grade ($\beta_{001} = 1.61$, $p < 0.001$) and CI ($\beta_{003} = 0.05$, $p = 0.001$) had a significant relationship with average knowledge score at post-test. Overall, teachers of advanced grades and teachers who were rated to administer the lessons with a higher level of accuracy (CI) produced students with better scores at post-test as measured by the WIST. DR ($\beta_{005} = -0.46$, $p = 0.357$), experience ($\beta_{002} = -0.001$, $p = 0.937$), and PI ($\beta_{004} = -0.41$, $p = 0.262$) did not have a significant impact on knowledge scores. Neither gender nor age were significant predictors at level 2. The growth rate remained significant ($\gamma_{100} = 0.79$, $p < 0.001$), with only grade significantly impacting the growth rate ($\beta_{101} = -0.25$, $p < 0.05$). Teacher-level characteristics significantly predicted knowledge gain over time with teachers of lower grades having students with a faster growth rate.

Discussion

This study sought to identify student- and teacher-level characteristics, including various aspects of treatment integrity, that contributed to knowledge gain for PK and elementary grade students participating in the Second Step CPU. Although previous research has suggested that this and similar programs can improve students' knowledge on child abuse prevention concepts (Nickerson et al. 2019; Zwi et al. 2007), most do not examine treatment integrity or maintenance of gains over time (Topping and Barron 2009). Student age and gender are related to baseline knowledge and growth; girls had higher post-test scores on the CSA knowledge measure, while older students scored higher on post-test than younger students on both measures. Over time, the gender effects become non-significant on the 6- and 12-month follow-up knowledge test. In addition, younger students had a faster growth rate on both measures. Although it appears that girls initially score higher than boys, these differences leveled out over time, indicating that both boys and girls benefit from this intervention across a 12-month period. In addition, although older students scored better at post-test, younger students learned faster across a 12-month period. Among all students, the grade taught also related to growth rate, with teachers of earlier grades having students who learned at a faster rate across the four time points. Findings highlighted the importance of CI (i.e., the extent to which teachers followed the manualized *Second Step* CPU) in improving student outcomes in terms of knowledge of CSA concepts and ability to recognize, refuse, and report unsafe situations in vignettes.

Student-Level Characteristics

Girls obtained higher post-test CSA knowledge scores, yet gender did not predict growth. Girls having higher scores at post-test is consistent with studies' findings that girls have greater knowledge of CSA prevention concepts (Chen and Chen 2005; Nickerson et al. 2019). These were no longer significant when including the 6- and 12-month follow-up, suggesting that both boys and girls benefit from the CPU. This is particularly relevant and promising given concerns that CSA prevention may disadvantage boys and fail to engage them (Scholes et al. 2014).

Older children scored higher on the post-test than younger children, with younger children making greater gains on both the CKAQ and the WIST. These results are consistent with findings from meta-analyses indicating higher effect sizes for younger (e.g., early elementary school students) than older students (Davis and Gidycz 2000; Rispens et al. 1997). In contrast to Rispens et al. (1997), who found that age effects disappeared at follow-up (with younger children not maintaining these gains), our study found that younger children made greater gains over time. There may be several reasons for this, including the repetition and reinforcement of the concepts from the first to second year of the program, as well as the behavioral skills practice that is part of the CPU. It is also possible that older students had some knowledge around these topics that the younger students did not, leading to lower initial scores and more room for improvement among younger students. The CPU was created to be developmentally appropriate, using activities such as music, puppets, and structured discussion to meet the needs of students across early education. Prior research demonstrates that these interventions may in fact be more impactful for younger students (Nickerson et al. 2019). Because younger students learned at a faster rate than their advanced peers, this suggests that students in PK and early elementary school can also benefit from these programs.

Teacher-Level Characteristics and Treatment Integrity

The primary aim of this study was to examine the extent to which different aspects of treatment integrity of the CPU related to children's outcomes. For students in grades 2–4, CI had a significant relationship with CKAQ post-test scores, indicating that greater accuracy and adherence to all aspects of the lessons (CI) by teachers was associated with greater

student knowledge. These findings are consistent with previous research that highlights the importance of program adherence (Dane and Schneider 1998; Gullan et al. 2009). In contrast to past research (Gullan et al. 2009) and contrary to our hypothesis, other dimensions of treatment integrity (i.e., PI, DR) did not relate to student outcomes. It is possible that there are other teacher-level variables that were not assessed that contribute significantly to student outcomes or that other factors attributing to quality of delivery may not have been captured. In addition, gender was still significant even when accounting for level-3 variables; girls had a significantly higher post-test score than boys, although the effect of gender was no longer significant when considering growth rate. Importantly, none of the teacher-level characteristics were related to the growth rate for the CKAQ. This means that there are important teacher-level characteristics contributing to the growth rate; however, they were not the variables assessed in the model. Some unassessed teacher-level characteristics that could contribute to student outcomes or growth rate could include self- and social-awareness, positive teacher-student relations, and overall acceptance of the intervention being implemented (Kim et al. 2019). In addition, work by Jennings and Greenberg (2009) highlights the importance of implementation self-monitoring, which could also influence variance. On the WIST (PK to 4th grade), greater CI (i.e., adherence to all aspects of the CPU) was associated with higher post-test scores. Teachers in advanced grades had students with higher scores at post-test. Although at post-test teachers of advanced grades had students with higher outcome scores, teachers of earlier grades had students who learned content faster across the 6-and 12-month follow-up, even when accounting for age at level 2. This demonstrates an additional component that may not be captured through the current variables, because these effects are due to the grade level of the teacher above and beyond the effects of student age. Future research may benefit from exploring what specifically contributes to differences among teachers of different grade levels that was not assessed in these models.

Teacher years of experience was not a significant predictor of student outcomes. In addition, although the level-3 growth rate model was significant for the CKAQ, no level-3 variables were associated with student knowledge gains over time. Only grade was significantly associated with knowledge gain as measured by the WIST. In addition, PI and DR were not associated with the intercept or growth rate in either model. It is possible that other validated constructs of effective teaching, such as emotional supports, classroom organization, and instructional supports (Pianta et al. 2008), could relate to these knowledge gains.

Limitations and Future Directions

Although this study addresses limitations of existing research on CSA prevention by using a large sample size, examining growth across multiple timepoints, and examining treatment integrity (Topping and Barron 2009), there are limitations. First, the measure of implementation fidelity, although based on previous research, was created by the researchers to be specific to this program and project. It is possible that using a validated measure of teaching, such as the *Classroom Assessment Scoring System* (CLASS; Pianta et al. 2008), would yield different results, particularly considering that there appeared to be teacher-level components that were not accounted for by the variables assessed in this study. In addition, observations of treatment integrity were based on the first year that the children received the CPU intervention; because the students changed classrooms and teachers but still received the intervention in the second year, it is possible that the integrity of implementation in that subsequent year could have impacted the results. Finally, the outcomes were all based on self-report knowledge measures as opposed to actual behavior or experience of CSA, which limits the extent to which we can make claims about the prevention program decreasing the risk of actual incidence of CSA.

In summary, PK to 4th grade students who received the CPU made gains in the knowledge of CSA concepts and skills over a 12-month follow-up period. Females generally scored higher than males on some outcomes measures immediately after the intervention, but gender did not impact knowledge gains over the four time periods, suggesting both boys and girls benefit from this prevention program. Children in later grades also had better knowledge scores at post-test, but an examination of growth rate revealed that younger students made greater gains on the CKAQ. This study shows support that treatment integrity is important, which Jennings and Greenberg (2009) suggest can be improved through self-awareness, social-awareness, and positive student relationships. Performance feedback and self-monitoring (i.e., monitoring and keeping track of one's implementation) have also been shown to improve treatment integrity (Noell et al. 2014). Although this study shows support for only CI impacting knowledge gain, it highlights the importance for assessing features of treatment integrity in intervention science. Future research should examine other classroom processes that may influence the efficacy of CSA prevention programs and lead to significant knowledge gains, as well as other underlying drivers of fidelity. Future research may also benefit from examining which aspects of the intervention specifically lead to knowledge gain (i.e., story and discussion, use of music, etc.).

Acknowledgments Special thanks to all of the research team that assisted with this project, including Kathleen Allen, Sunha Kim, Jennifer Livingston, Melissa Dudley, Jenine Tulledge, Samantha Kesselring, Timothy Parks, Peyton Schill, Kehinde Oladele, Nicole Castronovo, and Dylan Harrison, and to Megan Genovese and Hannah Grossman for contributions to the manuscript. We are grateful to the schools, parents, teachers, and students that supported this work.

Funding information This study was funded by the Committee for Children (no grant number through this organization).

Compliance with Ethical Standards

Conflict of Interest Amanda Nickerson, Ph.D., has received a research grant from the Committee for Children to fund this project. Margaret Manges, M.Ed., has received per diem for travel related to this project.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee (the University at Buffalo's Institutional Review Board; study number 00001263) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Passive informed consent was obtained from all parents, and informed assent was obtained from all students included in this study. Informed consent was obtained from all teacher participants included in the study.

Disclaimer The content is the responsibility of the authors and may not necessarily reflect the views of Committee for Children.

References

Allen, K. P., Livingston, J. A., & Nickerson, A. B. (2019). Child sexual abuse prevention education: A qualitative study of teachers' experiences implementing the *Second Step* Child Protection Unit. *American Journal of Sexuality Education*, 1–28. https://doi.org/10.1080/15546128/2019.1687382.

Ardoin, S. P., Binder, K. S., Foster, T. E., & Zawoyski, A. M. (2016). Repeated versus wide reading: A randomized control design study examining the impact of fluency interventions on underlying reading behavior. *Journal of School Psychology, 59*, 13–38. https://doi.org/10.1016/j.jsp.2016.09.002.

Briere, J., & Elliott, D. M. (2003). Prevalence and psychological sequelae of self-reported childhood physical and sexual abuse in a general population sample of men and women. *Child Abuse & Neglect, 27*, 1205–1222. https://doi.org/10.1016/j.chiabu.2003.09.008.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Sage Publications, Inc.

Chen, J. Q., & Chen, D. G. (2005). Awareness of child sexual abuse prevention education among parents of grade 3 elementary school pupils in Fuxin City, China. *Health Education Research, 20*, 540–547. https://doi.org/10.1093/her/cyh012.

Collier-Meek, M. A., Fallon, L. M., & Gould, K. (2018). How are treatment integrity data assessed? Reviewing the performance feedback literature. *School Psychology Quarterly*. https://doi.org/10.1037/spq0000239.

Committee for Children. (2014). *Second step child protection unit*. Seattle, WA: Author.

Daigneault, I., Vezina-Gagnon, P., Bourgeois, C., Esposito, T., & Hebert, M. (2017). Physical and mental health of children with substantiated sexual abuse: Gender comparisons from a matched-control cohort study. *Child Abuse & Neglect, 66*, 155–165. https://doi.org/10.1016/j.chiabu.2017.02.038.

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

Davis, M. K., & Gidycz, C. A. (2000). Child sexual abuse prevention programs: A meta-analysis. *Journal of Clinical Child Psychology, 29*, 257–265. https://doi.org/10.1207/S15374424jccp2902_11.

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*, 327. https://doi.org/10.1007/s10464-008-9165-0.

Finkelhor, D., Shattuck, A., Turner, H. A., & Hamby, S. L. (2014). The lifetime prevalence of child sexual abuse and sexual assault assessed in late adolescence. *Journal of Adolescent Health, 55*, 329–333.

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *A review and synthesis of the literature related to implementation of programs and practices*. Tampa, FL: Florida Mental Health Institute, National Implementation Research Network.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*, 79–88. https://doi.org/10.1016/j.cpr.2010.09.007.

Gullan, R. L., Feinberg, B. E., Freedman, M. A., Jawad, A., & Leff, S. S. (2009). Using participatory action research to design an intervention integrity system in the urban schools. *School Mental Health, 1*, 118–130. https://doi.org/10.1007/s12310-009-9006-9.

Hagermoser Sanetti, L. M., & Fallon, L. M. (2011). Treatment integrity assessment: How estimates of adherence, quality, and exposure influence interpretation of implementation. *Journal of Educational and Psychological Consultation, 21*, 209–232.

Irish, L., Kobayashi, I., & Delahanty, D. L. (2010). Long-term physical health consequences of childhood sexual abuse: A meta-analytic review. *Journal of Pediatric Psychology, 35*, 450–461. https://doi.org/10.1093/jpepsy/jsp118.

Jennings, P. A., & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79*, 49–525. https://doi.org/10.3102/0034654308325693.

Katerndalh, D. A., Burge, S. K., Kellogg, N. D., & Parra, J. M. (2005). Difference in childhood sexual abuse experience between adult Hispanic and Anglo women in primary care setting. *Journal of Child Sexual Abuse, 14*, 85–95. https://doi.org/10.1300/J070v14n02_05.

Kenny, M. C., & Wurtele, S. K. (2012). Preventing childhood sexual abuse: An ecological perspective. *Journal of Child Sexual Abuse, 21*, 361–367. https://doi.org/10.1080/10538712.2012.675567.

Kim, S., Nickerson, A. B., Livingston, J., Dudley, M., Manges, M., Tulledge, J., & Allen, K. (2019). *Teacher outcomes from Second Step Child Protection Unit: Moderating roles of preparedness and treatment acceptability*. *Journal of Child Sexual Abuse*. Online first publication. https://doi.org/10.1080/10538712.2019.1620397.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Mandell, D. S., Stahmer, A. C., Shin, S., Xie, M., Reisinger, E., & Marcus, S. C. (2013). The role of treatment fidelity on outcomes during a randomized field trial of an autism intervention. *Autism: The International Journal of Research and Practice, 17*, 281–295. https://doi.org/10.1177/1362361312473666.

Marquez-Flores, M. M., Marquez-Hernandez, V. V., & Granados-Gamez, G. (2016). Teachers' knowledge and beliefs about child sexual abuse. *Journal of Child Sexual Abuse, 25*, 538–555. https://doi.org/10.1080/10538712.2016.1189474.

Mellard, D. (2010). *Fidelity of implementation within a Response to Intervention (RtI) framework*. https://webnew.ped.state.nm.us/wp-content/uploads/2018/03/Fidelity-of-Implementation-guidev5-1.pdf

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation.

American Journal of Evaluation, 24, 315–340. https://doi.org/10.1177/109821400302400303.

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: Impact of intraclass correlation and sample size on type-1 error. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00074.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research, 39*, 374–396. https://doi.org/10.1007/s11414-012-9295-x.

Nickerson, A. B., Tulledge, J., Manges, M., Kesselring, S., Parks, T., Livingston, J. A., Dudley, M. (2019). Randomized controlled trial of the Child Protection Unit: Grade and gender as moderators of CSA prevention concepts in elementary students. *Child Abuse & Neglect, 96,* 1–12. https://doi.org/10.1016/j.chiabu.2019.104101

Noell, G. H., Gansle, K. A., Mevers, J. L., Knox, R. M., Mintz, J. C., & Dahir, A. (2014). Improving treatment plan implementation in schools: A meta-analysis of single subject design studies. *Journal of Behavioral Education, 23*, 168–191. https://doi.org/10.1007/s10864-013-9177-1.

Paolucci, E. O., Genuis, M. L., & Violato, C. (2001). A meta-analysis of the published research on the effects of child sexual abuse. *The Journal of Psychology, 135*, 17–30. https://doi.org/10.1080/00223980109603677.

Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice, 18*, 148–153. https://doi.org/10.1111/j.1468-2850.2011.01246.x.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12*, 365–383. https://doi.org/10.1093/clipsy.bpi045.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS)*. Baltimore: Brookes.

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

Rispens, J., Aleman, A., & Goudena, P. (1997). Prevention of child sexual abuse victimization: A meta-analysis of school programs. *Child Abuse and Neglect, 21*, 975–987. https://doi.org/10.1016/s0145-2134(97)00058-6.

Scholes, L., Jones, C., & Nagel, M. (2014). Boys and CSA prevention: Issues surrounding gender and approaches for prevention. *Australian Journal of Teacher Education, 39*, 1–15. https://doi.org/10.14221/ajte.2014v39n11.1.

Southerland, D. G., Farmer, E. M., Murray, M. E., Stambaugh, L. F., & Rosenberg, R. D. (2018). Measuring fidelity of empirically-supported treatment foster care: Preliminary psychometrics of the together facing the challenge—Fidelity of implementation test (TFTC-FIT). *Child & Family Social Work, 23*, 273–280.

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Education, 47*, 3–13. https://doi.org/10.1177/0022466911419516.

Topping, K. J., & Barron, I. G. (2009). School-based child sexual abuse prevention programs: A review of effectiveness. *Review of Educational Research, 79*, 431–463. https://doi.org/10.3102/0034654308325582.

Tutty, L. M. (1997). Child sexual abuse prevention programs: Evaluating "Who Do You Tell.". *Child Abuse & Neglect, 21*, 869–881. https://doi.org/10.1016/S0145-2134(97)00048-3.

Wurtele, S. K., Kast, L. C., & Kondrick, P. A. (1988). *Development of an instrument to evaluate sexual abuse prevention programs*. Atlanta, GA: Paper presented at the annual meeting of the American Psychological Association.

Wurtele, S. K., Kast, L. C., Miller-Perrin, C. L., & Kondrick, P. A. (1989). Comparison of programs for teaching personal safety skills to preschoolers. *Journal of Consulting and Clinical Psychology, 57*, 505–511. https://doi.org/10.1037//0022-006X.57.4.505.

Zwi, K., Woolfenden, S., Wheeler, D. M., O'Brien, T., Tait, P., & Williams, K. J. (2007). School-based education programmes for the prevention of child sexual abuse. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD004380.pub2.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.