REP
Resilience Education Program

Technical Report: Development and Evaluation of the Resilience Education Program (REP) Integrity Protocol

Stephen Kilgus, Alexandra Barber, Juliet Ye, & Katie Eklund

University of Wisconsin Madison

INTRODUCTION

Treatment integrity represents the extent to which an intervention is applied in accordance with its intended plan for implementation (Sanetti & Kratochwill, 2009). The examination of treatment integrity has grown increasingly common within research in recognition of its capacity to support decision making in many contexts (Gould et al., 2019). Within applied settings (e.g., school or clinic), treatment integrity evidence can strengthen the confidence one places in evaluations of a student's response to intervention (Kilgus et al., 2014). That is, if the student is responsive, practitioners can be more confident observed change is a result of the intervention. If non-responsive, practitioners can be assured it was not the lack of appropriate intervention implementation. Within research contexts, treatment integrity evidence can promote the validity of claims made regarding the efficacy or effectiveness of an intervention relative to a comparison condition (e.g., business-as-usual treatment; Peterson et al., 1982).

Modern definitions of treatment integrity decompose the construct into multiple dimensions of implementation (Sanetti & Kratochwill, 2009). Three of these dimensions have been more frequently examined within the literature (see Sanetti & Fallon, 2011).

1. ***Adherence*** represents the extent to which key steps of an intervention are implemented as intended. This is typically documented using checklists inclusive of these steps, with resulting scores representing the percent of steps correctly implemented.
2. ***Quality*** is assessed through examination of additional characteristics that make for more comprehensive intervention implementation. These characteristics are commonly evaluated using checklists inclusive of quality implementation steps. Rating scales are also appropriate, particularly when examining more subjective aspects of implementation that likely exist along a continuum (e.g., warmth, empathy, and positivity).
3. ***Exposure*** represents the extent to which intervention implementation followed its intended schedule. Exposure can be examined in terms of an overall intervention, as well as its major components. It can also be evaluated in terms of the number of opportunities for implementation (e.g., percent of intended implementation weeks), as well as the duration of implementation per each opportunity (e.g., percent of intended implementation minutes).

There are several broader methodologies through which treatment integrity can be assessed, including implementer self-report, permanent product review, and systematic direct observation (SDO). Self-report involves intervention implementers, such as teachers or mental health professionals, reporting on whether they delivered each

intervention step either during or following implementation. SDO involves third-party individuals (e.g., school psychologists) conducting in-vivo observations of intervention implementation. Finally, permanent product review involves the examination of materials that naturally result from implementation to determine whether key intervention steps were present.

Studies have revealed that implementers consistently overestimated the integrity with which they implement interventions (Noell, 2008); as a result, self-report is not recommended for use in evaluating treatment integrity (Sanetti et al., 2009). Research has also suggested that both SDO and permanent product review can yield valuable and unique data, and that each can serve different roles within a multi-method approach to treatment integrity assessment; with that said, in many cases, SDO data are likely to be more representative of actual treatment integrity and predictive of student response to intervention (Sanetti & Collier-Meek, 2014).

Though research has yielded broad and general conclusions about the utility of various treatment integrity assessment methods, it is rare for researchers to evaluate the psychometric defensibility of specific tools within these broader method categories. This is unfortunate, as one cannot assume that every SDO tool or permanent product review approach will yield consistently valuable and defensible data (Sanetti & Kratochwill, 2009). Accordingly, the purpose of this study was to evaluate the psychometric defensibility of a treatment integrity tool specific to the Resilience Education Program (REP), a Tier 2 targeted intervention for students exhibiting early signs of internalizing concerns (e.g., depression and anxiety). This measure, referred to as the **REP Integrity Protocol**, represents an SDO tool designed to assess the level of implementation adherence and quality for two major REP components: (1) small-group cognitive-behavioral instruction (CBI) and (2) Check In/Check Out (CICO). We posed a single research question: to what extent do REP Integrity Protocol scores demonstrate inter-observer reliability?

METHOD

Participants

School psychology graduate students (n = 9) participated in this study. Inclusion criteria included A) obtaining a passing grade in a course on behavioral, social, and emotional assessment, and B) attendance at a one-hour training session. All participants were required to complete all study components to be included in this study.

Measures

While viewing the videos, participants completed two different fidelity checklists that comprised the REP Integrity Protocol (Kilgus & Eklund, 2020), one for REP's CICO procedures and another for REP's CBI procedures. These checklists are described below:

REP CICO Fidelity Check. The REP CICO Fidelity Check is a 20-item checklist used to evaluate CICO fidelity via the examination of implementation adherence and quality

among teachers and mentors involved in CICO. Observers note whether teachers and mentors complete each step (***adherence***). Additionally, observers indicate whether certain implementation steps featured specific quality indicators (***quality***). These quality indicators include appropriate tone and nonverbal behavior, smooth/automatic, specific, and responsive. Each CICO element that is adhered to receives one point and each quality indicator observed also receives one point. Implementation and quality point totals are generated by summating all points earned for each fidelity category.

REP CBI Fidelity Check. The REP CBI Fidelity Check is a 12-item checklist used to evaluate CBI facilitators' implementation fidelity when conducting CBI group sessions. Observers indicate whether facilitators adhere to each necessary CBI step and note whether steps also featured quality indicators (appropriate tone and nonverbal behavior, smooth/automatic, specific, and responsive). Like the CICO Fidelity Check, each CBI step adhered to earns one point and each quality indicator observed also earns a point. Implementation and quality point totals are calculated on the CBI Fidelity Check by summating points for both areas.

Materials

Participants viewed four videos during this study. All videos featured the same actors who were all School Psychology graduate students. Videos were filmed by a graduate research assistant in an office space located on a university campus. Video length ranged from six and a half to eight minutes in length. In all four videos, the graduate student actors followed scripted lines and actions provided to them in advance of the recording session. These scripts all included the same characters (e.g., student, mentor) and had the same basic plot elements. Each script briefly demonstrated each of the three main components of CICO (check-in with mentor, teacher check-in, and check-out with mentor), and an abbreviated CBI lesson, but had slight variations based on condition.

The four video conditions were: (1) high quality/high adherence, (2) low quality/high adherence, (3) high quality/low adherence, and (4) low quality/low adherence. Scripts were systematically edited such that core components of the video remained constant and only specific quality and adherence indicators were adjusted. For example, in the two videos demonstrating low-quality fidelity, actors portraying teachers, mentors, and CBI facilitators were instructed to use a less expressive voice and give the student general feedback, while in the high-quality videos, the adult characters interacted with the student with enthusiasm and provided specific feedback. Differences in quality and adherence across videos allowed researchers to examine whether observers could adequately detect both the presence and absence of adherence and quality indicators in their observations.

Procedure

Participants were recruited through email. All eligible students from one School Psychology graduate program were emailed and asked to participate. All consenting participants then attended a participant training before viewing videos and completing fidelity checklists.

Participant Training. Participants attended a one-hour training led by the lead researcher and one project assistant. The training was held virtually using the Zoom platform. For most of the training, the lead researcher walked participants through an overview of REP, including detailed descriptions of the steps involved in REP's core treatment components. Next, they were led through a discussion of their role in the REP study. They learned about the adherence and quality indicators of fidelity used in REP and were introduced to the protocols that they would be using to evaluate the fidelity videos. Additionally, they were provided with a participant instructions flier which outlined their responsibilities as participants as well as a timeline for completing all study-related activities.

Data Collection. Following the training, participants were mailed paper copies of both fidelity checks. They were also emailed their participant number, their viewing order, and a link to access all four video files online. As the participants viewed the videos in their assigned order, they were instructed to complete the fidelity checklists by hand. Once they viewed all four videos, they were instructed to transfer their answers from the fidelity checklist forms into an online survey platform. Participants were specifically instructed to view each video just one time and were asked to avoid rewinding or re-watching videos at any point.

RESULTS

Participants' ratings on both quality and adherence indicators were aggregated. The percentage of affirmative ratings (e.g., answers of "Y" indicating yes) out of the total possible ratings was calculated. This resulted in mean percentage scores for total quality, total adherence, CICO quality, CICO adherence, CBI quality, and CBI adherence. All mean percentage scores are presented in Table 1.

Table 1. ***Mean Percentage of Implementation and Quality***

Condition	Total QUAL Mean	CICO QUAL Mean	CBI QUAL Mean	Total ADH Mean	CICO ADH Mean	CBI ADH Mean
High quality/ high adherence	55.72	73.06	40.48	59.61	67.83	45.46
High quality/ low adherence	41.38	57.83	26.52	46.67	54.40	33.36
Low quality/ high adherence	15.61	28.06	4.29	52.77	62.57	35.67
Low quality/ low adherence	5.57	8.37	3.03	41.49	47.37	32.34
Note: QUAL = Quality and ADH = Adherence						

A series of Fleiss' kappa coefficients were calculated in examining inter-observer reliability within each of the four videos. The first set of coefficients were considered omnibus indicators of reliability across both REP components (i.e., CBI and CICO) and

type of treatment fidelity (i.e., adherence and quality). Two of the kappa coefficients exceeded the What Works Clearinghouse (WWC) threshold of .50 for acceptability reliability, with the lower bound of their corresponding confidence intervals also exceeding this threshold. Two other coefficients closely approximated this threshold, with kappa equal to .49 for each and upper bounds of their confidence intervals exceeding this threshold. All kappa coefficients were statistically significant at the $p <$ .001 level. Kappa statistics and associated 95% confidence intervals are reported below for each video:

Table 2. ***Omnibus Kappa Coefficients across Videos***

•**Low Adherence/High Quality** κ = .49 (CI-95 = .47, .51)	•**High Adherence/Low Quality** κ = .57 (CI-95 = .55, .60)
•**Low Adherence/Low Quality** κ = .54 (CI-95 = .52, .56)	•**High Adherence/High Quality** κ = .49 (CI-95 = .46, .51)

Four additional kappa coefficients were calculated within each video clip. The first two coefficients were specific to inter-observer reliability for adherence and quality codes (respectively) collapsed across both CBI and CICO. The second two coefficients were specific to CBI and CICO codes (respectively) collapsed across adherence and quality. Each of these kappa coefficients and corresponding 95% confidence intervals are reported below:

Table 3. ***Kappa Coefficients across REP Components and Fidelity Dimensions***

Codes	Low ADH/ High QUAL	Low ADH/ Low QUAL	High ADH/ Low QUAL	High ADH/ High QUAL
Adherence	.41 (.38, .43)	.51 (.45, .57)	.58 (.52, .64)	.52 (.46, .58)
Quality	.52 (.46, .57)	.45 (.42, .48)	.48 (.45, .51)	.41 (.38, .44)
CICO	.44 (.41, .47)	.44 (.41, .47)	.54 (.51, .57)	.43 (.40, .46)
CBI	.47 (.43, .50)	.62 (.58, .65)	.54 (.51, .57)	.45 (.42, .49)
Note: QUAL = Quality and ADH = Adherence				

DISCUSSION

This study examined the inter-observer reliability of REP Integrity Protocol scores, which graduate students generated while viewing videos of REP implementation. Participants rated the extent to which interventionists adhered to REP implementation steps, as well as the quality of their implementation. Data were collected for both CICO and CBI components of REP. Across the four video conditions (High Adherence/High Quality, High Adherence/Low Quality, Low Adherence/High Quality, and Low Adherence/Low Quality), omnibus kappa coefficients in both the Low Adherence/Low Quality and High Adherence/Low Quality exceeded the WWC threshold of 0.50, while the other conditions had kappa coefficient values that approached 0.50. This suggests that the participants were able to achieve sufficient reliability following minimal training. The

differences in reliability between conditions were minimal; however, Low Quality conditions achieved slightly higher reliability. This could indicate that it was easier for participants to observe the absence of quality in implementation as opposed to the presence of it. This has implications for revisions to training procedures for REP Integrity Protocol users.

Kappa coefficients were also calculated within each video clip to compare the reliability of coding for different aspects of fidelity (i.e., adherence and quality) as well as the different components of REP (i.e., CBI and CICO). Though differences were small, there was stronger reliability in adherence codes compared to quality codes, and in CBI codes compared with CICO codes. Although inter-observer reliability reached the threshold for acceptability in many circumstances, it did not in others. This suggests the need to consider more comprehensive training approaches moving forward, with particular attention given to preparing observers to code CICO implementation and quality of implementation more generally.

Overall, the REP Integrity Protocol proved to be an acceptable tool for observing implementation fidelity. As such, this supports further utilization of the REP Integrity Protocol both in practice and in future research endeavors. As SDO data are likely to provide a more accurate and generalizable estimate of actual treatment integrity relative to other methods (Sanetti & Collier-Meek, 2014), scores collected from the REP Integrity Protocol can be used to pinpoint areas where treatment integrity of CBI and CICO components may be strengthened. This would increase confidence that the intervention has been implemented as intended, therefore increasing one's in the internal validity of subsequent decisions regarding whether students have responded to intervention or not (Kilgus et al., 2014; Peterson et al., 1982). As such, tools such as the REP Integrity Protocol are important instruments to further understand and include when implementing and evaluating interventions.

REFERENCES

1. Gould, K. M., Collier-Meek, M., DeFouw, E. R., Silva, M., & Kleinert, W. (2019). A systematic review of treatment integrity assessment from 2004 to 2014: Examining behavioral interventions for students with autism spectrum disorder. *Contemporary School Psychology*, *23*(3), 220-230. https://doi.org/10.1007/s40688-019-00233-4
2. Kilgus, S. P., Collier-Meek, M. A., Johnson, A. H., & Jaffery, R. (2014). Applied empiricism: Ensuring the validity of causal response to intervention decisions. *Contemporary School Psychology, 18*, 1-12. https://doi.org/10.1007/s40688-013-0009-z
3. Kilgus, S. P., & Eklund, K. (2020). *Resilience Education Program (REP) Integrity Protocol*. Wisconsin Center for Education Research. https://smhcollaborative.org/rep-materials/
4. Noell, G. H. (2008). Research examining the relationships among consultation process, treatment integrity, and outcomes. In W. P. Erchul & S. M. Sheridan (Eds.),

Handbook of research in school consultation: Empirical foundations for the field (pp. 315–334). Erlbaum.

5. Peterson, L., Homer, A., & Wonderlich, S. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15*(4), 477–492. https://doi.org/10.1901/jaba.1982.15-477
6. Sanetti, L. M. H., & 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. https://doi.org/10.1080/10474412.2011.595163
7. Sanetti, L. M. H., Chafouleas, S. M., Christ, T. J., & Gritter, K. L. (2009). Extending use of Direct Behavior Rating beyond student assessment: Applications to treatment integrity assessment within a multi-tier model of school-based intervention delivery. *Assessment for Effective Intervention, 34*(4), 251–258. https://doi.org/10.1177/1534508409332788
8. Sanetti, L. M. H., & Collier-Meek, M. A. (2014). Increasing the rigor of procedural fidelity assessment: An empirical comparison of direct observation and permanent product review methods. *Journal of Behavioral Education, 23*(1), 60-88. https://doi.org/10.1007/s10864-013-9179-z
9. Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review, 38*(4), 445–459.

Learn more about the School Mental Health Collaborative

School Mental Health COLLABORATIVE

Co-Directors: Evan Dart, PhD, Katie Eklund, PhD, Andy Garbacz, Ph., Stephen Kilgus, Ph., Shannon Suldo, PhD, and Nate von der Embse, PhD

Mission: The purpose of the School Mental Health Collaborative (SMHC) is to conduct research that informs policy and practice related to the promotion of social-emotional and behavioral success of all students. SMHC scholars generate tools, resources, and guidance that help educators and parents promote the mental health of children and adolescents.

To Cite this Technical Report: Kilgus, S., Barber, A., Ye, J., & Eklund, K. (April, 2022). *Development and evaluation of the Resilience Education Program (REP) Integrity Protocol (Technical Report No. 2022-1)*. https://smhcollaborative.org/rep-activity-1-2-tech-report-final/

The research reported here was supported by the **Institute of Education Sciences (IES)**, U.S. Department of Education, through Grant R324A190129 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.