11-16-2017

Generalizability of Multiple Measures of Treatment Integrity: An Empirical Replication

Elizabeth Wilson
Louisiana State University and Agricultural and Mechanical College, ewils24@lsu.edu

GENERALIZABILITY OF MULTIPLE MEASURES OF TREATMENT INTEGRITY: AN EMPIRICAL REPLICATION

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Arts

in

The Department of Psychology

by
Elizabeth Wilson
B.A., Louisiana State University, 2012
M.S., University of Texas at Dallas, 2014
December 2017

TABLE OF CONTENTS

Abstract

Treatment integrity is essential for the implementation of interventions in schools as it determines the accuracy or consistency with which different components of a treatment are implemented. There are no current standards regarding the best practices in treatment integrity measurement; however, higher integrity is associated with enhanced student outcomes. At present, there is no database providing information on treatment integrity for practitioners, researchers, and policy-makers to reference for choosing an appropriate level of treatment integrity needed for certain interventions for certain problems. Consequently, there is a need to establish convergent validity among different methods of treatment integrity measurement using multiple evidence-based interventions in order to guide best practices. The current study attempted to replicate and expand the finding that the direct observation method yields the most reliable treatment integrity data the most quickly, followed by self-report, when using an evidence-based intervention (Gresham, Dart, & Collins, 2017). For this study, researchers empirically replicated the methods of Gresham and colleagues' work to examine two of the three measures of treatment integrity, direct observation and self-report, for six teachers' implementation of the response cards intervention.

Keywords: treatment integrity, intervention adherence, Generalizability Theory

Generalizability of Multiple Measures of Treatment Integrity: A Conceptual Replication

Methods of treatment integrity measurement are a necessity in implementation science, as they assess the accuracy in which an intervention is implemented (Gresham, 1989). With information collected from measures of integrity, scientists can ensure that the changes in the dependent variable are attributable to the intervention and not to outside variables. The concept of treatment integrity is critical in all fields, not just psychology, that are involved in providing intervention or treatment to individuals. Different terms are used to explain integrity within the field of psychology. In particular, school psychology refers to this concept as "treatment integrity" or "implementation plan integrity" (Gresham, 1989). Although there are various names of the concept, they all share a similar goal: to ensure that desired changes in an individual's functioning or behavior are due to the systematic and accurate implementation of treatment and not extraneous variables (Gresham, 2009). In previous research, higher levels of treatment integrity were correlated with better intervention outcomes (Durlak & DuPre, 2008; Fiske, 2008; Sanetti & Kratochwill, 2009). High treatment integrity is vital in behavioral interventions implemented in schools because of the many factors that can affect the accuracy of intervention implementation and the resulting behavior change, such as teacher characteristics, time and resources, difficulty level, and number of steps involved in the implementation (McIntyre, Gresham, DiGennaro, & Reed, 2007). Most of the time, school psychologists are having to train and ultimately change the behavior of teachers who have demanding schedules and have the ability directly or indirectly to refuse to participate or implement the intervention correctly (Noell, Gansle, Mevers, Knox, Mintz, & Dahir, 2014).

In practice, a functional relationship is assumed between the delivery of treatment, most often done by third parties such as teachers and parents, and the outcome of the intervention;

however, the factors mentioned above can change the intervention and the way it is implemented in ways unknown to the researcher. This can lead to researchers and practitioners making assumptions that certain treatments are working for certain issues, when in reality there are extraneous variables affecting the measureable outcomes. This can also have the reverse effect, in which a treatment appears to fail to produce the desired outcomes and the researcher assumes the treatment is ineffective, when the failure to evoke the desired change could be due to outside factors (Gresham, 2009). Treatment integrity is not synonymous with intervention effectiveness, which is defined as how strong the treatment is in producing the desired outcome. If a treatment is implemented with low integrity and does not produce the desired changes, it could be that it was implemented differently than originally planned, resulting in an intervention that is not representative of what the researchers intended (Perepletchicova, Hilt, Chereji, & Kazdin, 2009).

Treatment integrity has been conceptualized as the degree to which a treatment is implemented in the way that it was intended, however not all treatments are as resistant to outside variables at the same level. Some treatments are stronger than others and are more resistant to poor implementation, therefore these treatments would rely on a high level of treatment integrity to be effective. Whereas weaker treatments can be implemented flawlessly and still have minimal effects on the variable of interest (Gresham, 2009). Thus, different treatments require different levels of treatment integrity in order for the treatment to produce significant changes.

Treatment Integrity and Policy

In recent years, there has been more concern in the field of school psychology regarding the concept of treatment integrity and its importance in intervention implementation; however, there has been little attention given to the empirical measurement of treatment integrity. The

Institute of Educational Sciences (IES) within the U.S. Department of Education requires that for every intervention, there should be a measure of treatment integrity, but it does not specify or provide any direction as to what levels of treatment integrity are required for specific interventions. IES is moving in the right direction by addressing and requiring treatment integrity, but they do not provide any guidelines as to best practices for measuring treatment integrity. There is no guidance for practitioners regarding which methods of treatment integrity measurement should be used for specific interventions or how many assessments of integrity are needed over what period of time.

The reason for the lack of information regarding these critical issues in treatment integrity is that there is no database providing information on treatment integrity for practitioners, researchers, and policy-makers to reference for choosing an appropriate level of treatment integrity needed for certain interventions for certain problems. There are no current standards regarding the best practices in treatment integrity measurement. As discussed previously, different treatments for different problems may require different levels of treatment integrity to yield the desired change in the dependent variable. Researchers know the true value of an intervention before implementation because they have standardized the individual steps that interventionists should take when implementing the intervention. However, different methods of treatment integrity measurement may provide researchers with different reported levels or degrees of treatment integrity, based on how reliable these different measurements are at producing an accurate or "true value" of behavior (Gresham, Dart, & Collins, 2017).

Measurement of Treatment Integrity

Treatment integrity can be captured through various approaches. The researchers will be examining two of the most common methods of collecting treatment integrity, which are self-

report and direct observations (Fiske, 2008). Self-report methods rely on the agent of change, in this case the teacher, to report on their completion of the components of the intervention. Direct observation is the most commonly used method, in which, typically, a consultant will observe the teacher during the intervention utilizing a components checklist. After the observation, the consultant will calculate the percentage of components completed (Noell, Witt, Slider, & Connell, 2005).

Even though these methods are frequently relied upon to reflect accurate measures of treatment integrity, studies have shown that these assessments do not possess strong psychometric qualities (Sanetti & Kratochwill, 2009). For example, direct observations have been shown to cause reactivity in intervention agents (Sheridan et al., 2009) and may not reflect accurate treatment integrity over time due to observer drift, which is an unintended change in the accuracy of the observer's performance (Gresham et al., 2009). In addition, self-reports have been shown to display a severe upward bias, in which teachers will report themselves as completing the intervention or parts of the intervention correctly when in reality they did not (Noell et al., 2014).

Generalizability Theory

Convergent validity among different measurement methods for treatment integrity needs to be established across multiple evidence-based interventions in order to build a reliable database. Gresham (2009) proposes the idea of establishing a "treatment integrity effect norms" database, in which multiple researchers across multiple sites work to quantify "what levels of treatment integrity, measured by what methods, with what intervention procedures, produces what level of treatment integrity outcomes" (pp. 5). Gresham et al. (2017) is the first study to examine the dependability of different methods of treatment integrity of a widely known

evidence-based intervention, the Good Behavior Game, using Generalizability Theory methods. Generalizability Theory is a statistical framework that is used to evaluate the dependability or reliability of behavioral measurements (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). The researchers sought to begin the first of a series of studies that would assess multiple methods of treatment integrity assessment across multiple evidence-based interventions.

Gresham and colleagues assessed treatment integrity twice per week and used a .80 reliability criterion. They found that although all three methods were reliable, direct observation provided the most reliable measure of treatment integrity (i.e., 4 observations yielded reliable data), followed by permanent products (i.e., 5 observations provided a reliable measure of treatment integrity), and then self-reports (i.e., 7 observations yielded reliable data). The researchers also found low correlations between the three measures, with significantly lower correlations between self-report and the other two methods.

The previous study is the first to apply Generalizability Theory to the assessment of treatment integrity and enhances the existing research that the direct observation method is the standard in treatment integrity assessment when examining behavior in schools. The results are critical to the field as they guide practitioners as to which method to use when making important decisions in schools, and they are the first of many studies that will add to a database of treatment integrity effectiveness norms.

Purpose of the Present Study

Given the lack of data we have on measures of treatment integrity for various interventions, the purpose of the current study is to replicate the findings of Gresham et al. (2017), by extending the methods to another well-established, academic, evidence-based intervention, Response Cards (Christle & Schuster, 2003). The Response Card intervention was

selected because it has been shown to increase academic engagement, student participation in whole class activities, on-task behavior, and academic achievement when compared to usual strategies used by teachers as an attempt to increase these factors (i.e., hand raising, cold-calling, etc.) (Christle & Schuster, 2003). Gresham et al. (2017) examined the Good Behavior Game, which is a class-wide, behavioral intervention. Researchers were interested in the possibility of the positive results found with the Good Behavior Game, a behavioral intervention, generalizing to the Response Card Intervention, an academic intervention with related behavioral benefits. Researchers examined two methods of assessing treatment integrity, which are direct observation and self-report, when used to assess teachers' implementation of Response Cards. Researchers used Generalizability Theory to conduct a decision (D) study in order to determine how many observations each method of integrity requires so that it reaches a point at which it provides a reliable measure of treatment integrity. In other words, researchers investigated how efficient each measure of treatment integrity is when general education teachers implement the Response Card intervention. Additionally, researchers analyzed the correlation between the two measures of treatment integrity. It was hypothesized that the results of this study would align with the Gresham et al. (2017) study, with direct observation yielding the most reliable measure of teacher implementation followed by self-report.

Method

Participants and Setting

The primary participants in the study were six elementary school teachers from a low-income, high needs public school in East Baton Rouge Parish. To be included in the study, teachers had to primarily teach classes from third through fifth grade. Two fifth grade teachers were disqualified from the analyses due to missing data as a result of scheduling conflicts during the end of year testing period. All teachers were between the ages of 20 and 35, taught general education classes, and had less than 5 years of teaching experience. Teacher 1 was a Caucasian female who taught third grade, Teacher 2 was a Caucasian female who taught third grade, Teacher 3 was a Caucasian female who taught fourth grade, and Teacher 4 was a Hispanic female who taught fourth grade. Researchers sought teachers who operate with set blocks of direct instruction built in to their daily schedules. To participate in this study teachers had to be willing to be responsible for the implementation of the evidence-based, class-wide intervention (i.e., Response Cards) once every afternoon to increase academic engagement in their classrooms. The afternoon block was chosen as a way of keeping time of day consistent across teachers. Teachers decided during which afternoon block, or academic subject, to implement the intervention depending on when they perceive academic engagement to be lowest. All of the study procedures took place within each teacher's classroom.

Five graduate students in the school psychology doctoral program at Louisiana State University who had previous training in systematic direct observation, consultation with teachers, and classroom management strategies participated in the study. The graduate students were responsible for the initial consultation with the teachers and establishing the Response Card

intervention in the classrooms. They also trained the teachers on intervention implementation and collected treatment integrity twice per week throughout the duration of the study.

Measures

Two different measures of treatment integrity were analyzed – direct observation and self-report. These specific measures were chosen because they are found to be the most commonly used methods to measure treatment integrity in practice (Lane, Boccian, MacMillan, & Gresham, 2004). Each component of the Response Card intervention was operationalized so that these components could be standardized across both measures. The seven components are: (a) the teacher states to the class that it is time to respond using response cards, (b) the teacher reviews the rules of the intervention with the class, (c) the teacher presents a question to the class, (d) the teacher provides adequate time for students to write their answers on the response cards, (e) the teacher requests that students present their cards, (f) if less than 75% of students provide correct answers, the teacher instructs the class to correct their answers and provide a rationale for the answer; if 75% or more of the class answers correctly, the teacher reveals the correct answer to the class, (g) the teacher provides verbal praise for correct responses (Lambert, Cartledge, Heward, & Lo, 2006).

Direct observation. In order to assess the percentage of components completed by teachers during a given intervention period, an observation form (see Appendix A) was created in which the seven items directly match the seven components provided above. Observers circled either "yes" or "no" depending on whether they observed the teacher complete the component. All of the components are discrete, explicit events that could be easily identified by the observer. Treatment integrity was calculated by dividing the number of "yes" components by the total number of components listed and multiplying by 100 to yield a percentage of treatment integrity.

Self-report form. The self-report form (see Appendix B) is identical to the observation form, but was completed by the teacher immediately after every intervention period that the observer was present. The teachers circled either "yes" or "no" depending on whether they completed the component. Treatment integrity was calculated in the same way as direct observation, by dividing the number of "yes" components by the total number of components listed and multiplying by 100 to yield a percentage of treatment integrity.

Intervention materials. The response cards were created from half pages of card stock paper that were then laminated so that they would be durable. Each teacher was provided with enough response cards so that every student in the classroom had one. The same number of markers and socks, used as erasers, were provided to the teachers as well. The teachers were also given a self-report form by the observer each day that the teachers were observed.

Procedure

Before participants were recruited, the Institutional Review Board at Louisiana State University reviewed and approved the study methods and procedures. Administrative consent was obtained, which allowed the researchers to recruit in their respective schools. Teachers were recruited to participate in this study if they had reached out for consultation services regarding classroom management strategies in the past. Research assistants from Louisiana State University currently provide consultation services at the public school, so it is common for these services to be requested by teachers, especially in the elementary grades. Administrators also suggested teachers they thought would benefit from learning how to implement the intervention. The first teachers to request services were recruited and given a brief explanation of the study. Research assistants explained the procedures and timeline of the study, and obtained informed consent from the first six teachers to agree to participation.

Once a teacher provided consent to participate in the study, a meeting was scheduled with the teacher, where the primary researcher explained the Response Card intervention and helped to establish the intervention in the classroom. Teachers were asked to select an academic subject block in the afternoon (i.e., math, ELA, science, etc.) that they perceived to be problematic in terms of participation and academic engagement, so that the intervention would be implemented during that time period. The teachers were instructed to implement the Response Card intervention for 15 to 25 minutes each day during the selected academic block.

Research assistants trained teachers to implement the Response Card intervention, by explaining and modeling the intervention. Teachers were given all of the materials necessary to implement the intervention and were given an opportunity to go through the materials and ask any questions. During this initial training session, teachers were given a typed protocol (Appendix C), which included in depth instructions of what would happen on observation days as well as outlined steps of the intervention in greater detail than the self-report forms they would use to fill out during the observation days. These protocols were for the teachers to keep and refer back to if they had any questions.

Observation schedule. A randomized treatment integrity assessment schedule was created for each of the six teachers in order to collect a representative sample of intervention implementation. Teachers were required to implement the Respond Card intervention five times each week (i.e., once per afternoon during the designated block). For each of the five weeks, two of the five possible implementations per teacher were randomly selected for observation, resulting in 10 total observations per teacher – 60 observations overall. According to Ferguson, Briesch, Volpe, & Daniels (2012), over the course of one day, a 15-minute observation was found adequate to yield reliable data for making low stakes decisions. Observations for the

current study lasted anywhere from 15 to 25 minutes each, depending on how long it took the teacher to get started with the intervention and if there were disruptions during the intervention. During the observation, the primary researcher directly observed the intervention implementation for treatment integrity, and filled out the appropriate form to acquire an objective measure. Immediately after the intervention block, the teacher filled out a self-report integrity form; consequently, each random observation resulted in one assessment of treatment integrity per each of the different sources. For the purpose of necessary interobserver reliability, there were two graduate students present for 25% of each teacher's treatment integrity observations. Though graduate student research assistants had formal training in systematic direct observation, they also attended a training specific to this study led by the main researcher. All graduate research assistants were given a typed protocol (Appendix D), which outlined the observation schedule, procedure, materials needed, and general guidelines regarding observer duties for the direct observation period.

Results

The researchers conducted a G study to evaluate the generalizability of each method of treatment integrity. A fully crossed analytic design with two facets was created in which treatment integrity method (*m*: i.e., self-report and direct observation) was fully crossed with person, the object of measurement (*p*: i.e., teachers). Person was also fully crossed with occasion (*o:* i.e., observation), and occasion was fully crossed with method (see Figure 1 for conceptualization). In other words, teachers' implementation of the intervention was assessed using both of the treatment integrity measurement methods during each observation. The observation schedule was randomized so that each teacher was observed twice per week for five weeks, alternating among the five school days of the week. Furthermore, a follow-up D study was conducted to establish the G and Phi coefficients associated with each treatment integrity method when examining different numbers of observations. The D study determined the amount of observations needed for each method to achieve reliable treatment integrity scores when manipulating the most malleable facet (occasion) using a cutoff criterion of .80 (Briesch, Swaminathan, Welsh, & Chafouleas, 2014). In other words, researchers determined how many days each method must be used before producing a reliable treatment integrity score (> .80). All analyses were conducted using syntax specifically written for generalizability theory analyses in SPSS (Mushquash & O'Conner, 2006). Finally, a Pearson's correlation (*r*) matrix was constructed for each method and correlations between the results of each method were tested.

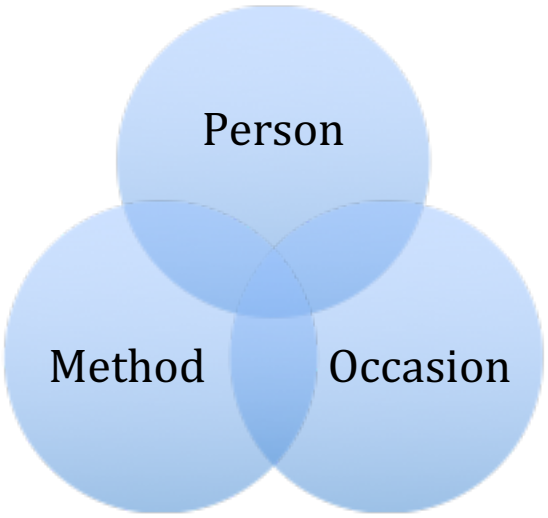


Figure 1. 2-facet, fully crossed analytic design: *m x p x o*. This figure demonstrates the analytic design of the current study. Teachers' implementation of the intervention will be assessed using both of the treatment integrity measurement methods during every observation.

Researchers hypothesized that results would show that the method of direct observation would produce the most reliable measure of treatment integrity followed by the self-report measure. Although previous research has shown that adequate reliability was not found when such a limited sampling of observations was conducted (i.e. twice per week) it was expected that, as in the Gresham et al (2017) study, the simplicity of the measurement systems (i.e. 7 components of teacher behavior per method over a 15 to 25-minute period during an academic block) versus the momentary time sampling used in previous studies, would yield reliable data (Hinze & Matthews, 2004). Finally, it was expected that correlations would be low between the two methods as previous literature suggests that teachers would be more likely to rate themselves as completing the components (Wickstrom, Jones, LaFleur, and Witt, 1998).

The results of the G study are presented in Table 1. Estimations of variance in treatment integrity scores associated with each facet and interactions between facets were compared to total variance of scores to examine the percentage of variance accounted for by each. Once

variance components were computed, it was possible to calculate a G coefficient and Phi coefficient signifying relative and absolute dependability of scores respectively, which determined if the methods produced sufficiently reliable estimates of treatment integrity when measured twice per week for five weeks. Overall, error variance was the single largest contributor to variance in treatment integrity rating scores accounting for 71.4% of the variance. Since Method (*m*: type of treatment integrity method: self-report vs. direct observations) is a fixed facet, the variance for each type of method could not be parsed, however, the Method facet accounted for 15.8% of the overall variance. In other words, the ratings scores varied by 15.8% depending on which method of treatment integrity was used to obtain those scores. The Occasion (*o*: observation day) facet accounted for 4.4% of the variance, meaning that scores varied depending on which observation (e.g., observation 1, observation 2, etc.) was examined. The Person (*p*: teacher) facet accounted for 2% of the variance, meaning that scores varied depending on individual teacher ratings. Finally, the interaction between the Person and Method facets accounted for 6.5% of the variance.

Table 1. Proportion of Variance for Each Facet

Facet	Proportion of Variance
Person (*p*)	2
Method (*m*)	15.8
Occasion (*o*)	4.4
Person x Day	6.5
Error	71.4

Once the variance components were computed it was possible to calculate a G coefficient and Phi coefficient, which represent the relative and absolute dependability of the scores generated by each assessment method, respectively. The G coefficient was higher for the direct observation method (.231) followed by self-report (.13). The Phi coefficient was also higher for the direct observation method (.119) followed by self-report (.064) (see Table 2). These coefficients suggest that neither method produces sufficiently reliable estimates of treatment integrity when measured twice per week for five weeks, regardless of the type of decision being made (i.e., relative or absolute).

Table 2. Dependability Coefficients from Each Assessment Method

	Direct Observation	Self-Report
G Coefficient	.231	.13
Phi Coefficient	.119	.064

Decision Studies

Because the G study suggested that the original measurement model did not produce sufficiently dependable assessment data across the two methods, a decision study was conducted to determine the point at which each method would produce a reliable assessment when manipulating the most malleable facet, Occasion (i.e., number of observations). Using .80 as a cutoff criterion for adequate dependability (e.g., Briesch, Swaminathan, Welsh, & Chafouleas, 2014), the relative dependability of each measure of treatment integrity only minimally increases when expanding the D study out to 100 observations, still assuming a rate of two observations

per teacher per week. At day 50, neither the direct observation nor self-report method would produce reliable treatment integrity assessments, with G coefficients at .341 and .206 respectively (see Table 3).

Table 3. G Coefficients for Each Method as a Function of Days of Assessment

	Days									
	5	**10**	15	20	25	30	35	40	45	50
Self-Report	.090	**.130**	.154	.169	.180	.188	.194	.198	.202	.206
Direct Observation	.164	**.231**	.267	.289	.305	.316	.324	.331	.337	.341

The bolded numbers indicate the G coefficients for the number of observations in the current study.

Correlation

Pearson's correlation (r = .051) was not significant between direct observation and self-report methods, as expected.

Discussion

The current study is an empirical replication of the Gresham et al (2017) study and results were expected to align accordingly. The purpose of the study was to determine the dependability of two different measures of treatment integrity when general education teachers implement the Response Card intervention. Repeated measurement involving direct observation and self-report allowed for comparison of two of the most common measures of treatment integrity in schools. Results indicated that neither method produced a reliable measure of treatment integrity, when teachers were observed twice per week for five weeks. Even when expanding the amount of observations out to 100 days, the G coefficients reached a ceiling of .381 for direct observation and .206 for self-report. The results are possibly due to the infrequency of observations. Teachers were typically observed twice per week for five weeks, however, taking into account absences and "make-up observations," observations were pushed even further apart. The reasoning for observing teachers only twice per week was to mimic a typical observation schedule used in schools, since twice per week is usually most feasible for school psychologists and other school professionals. Perhaps future studies would benefit from observing teachers for ten consecutive days or every other day so that those reliability estimates could be calculated. More frequent observations would give a clearer picture of teacher implementation since there would not be so much time between observations and opportunity for possible extraneous variables to have an effect.

As expected, there was not a significant correlation between the two methods. A better way to visualize the relationship between the two methods would be to look at the way in which the treatment integrity ratings for each method changed over the course of the ten observations for each teacher, and if the change was consistent in the same way. See figures 2 through 5 for a

representation of how the two treatment integrity methods changed between individual observations for each teacher. The relationship between the two methods does not change consistently over time, which is to be expected. Many of the teachers reported 100% integrity on each day of data collection, while direct observation ratings yielded more varied, and likely more accurate, ratings of treatment integrity. Even though neither method produced reliable measures of treatment integrity, these almost constant 100% treatment integrity ratings from teachers' self-reports supports the idea that this data is not an accurate estimate of treatment integrity. Additionally, the IOA calculated for this study came to 66%. One of the raters who participated in some of the direct observations also reported 100% integrity on each day of data collection, which indicates a need for more intensive, in vivo, rater training for future studies to ensure adequate understanding of the items and requirements of implementation up to a certain criterion. It could also be beneficial for future studies to include rater as a facet to examine how much variance comes from differences in individual raters.

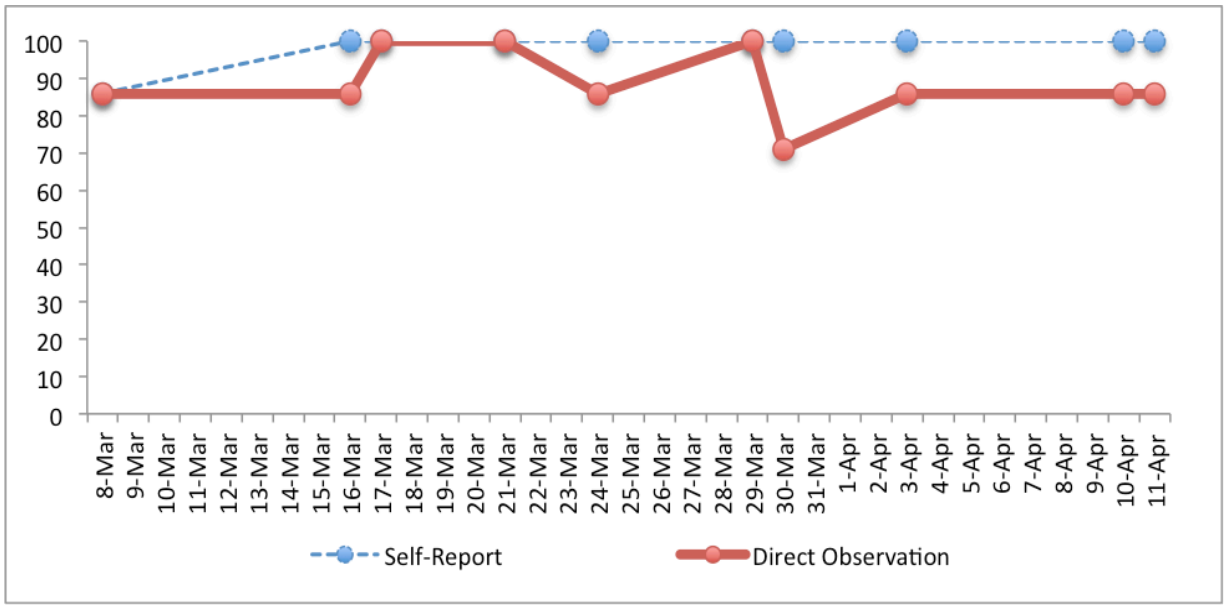


Figure 2. Self-report and direct observation ratings over the data collection period for Teacher 1.

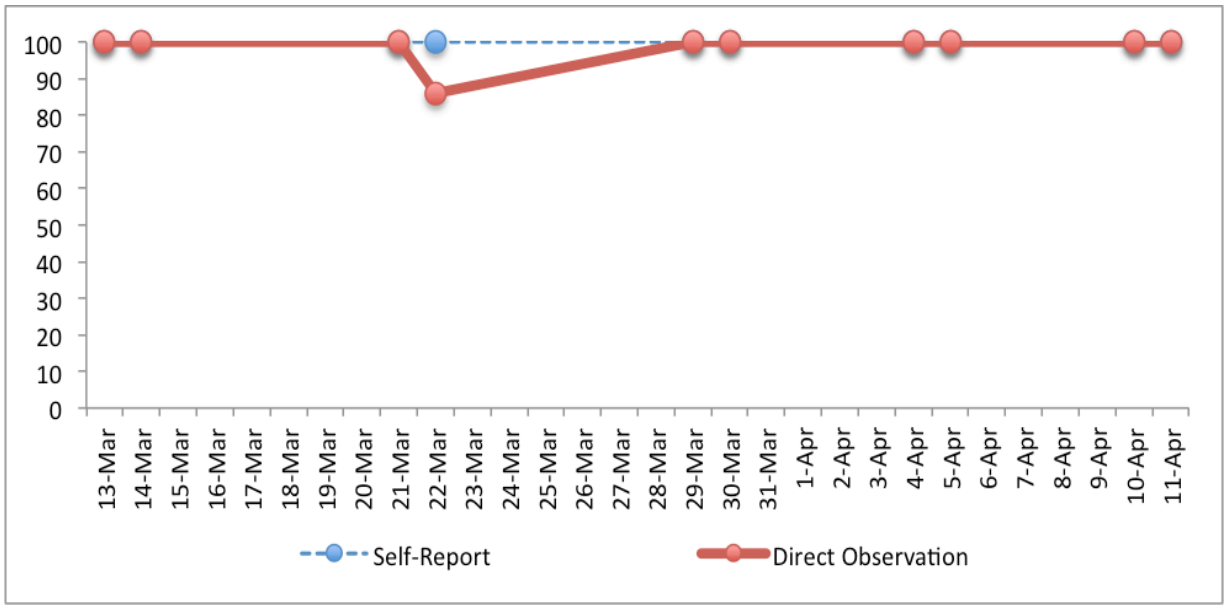


Figure 3. Self-report and direct observation ratings over the data collection period for Teacher 2.

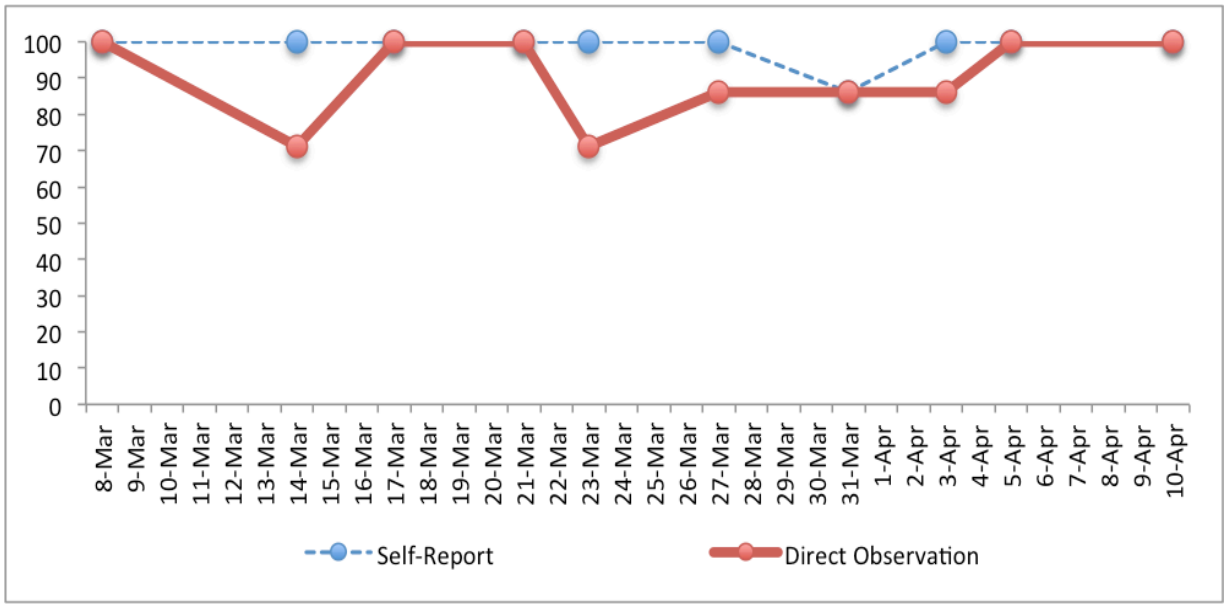


Figure 4. Self-report and direct observation ratings over the data collection period for Teacher 3.

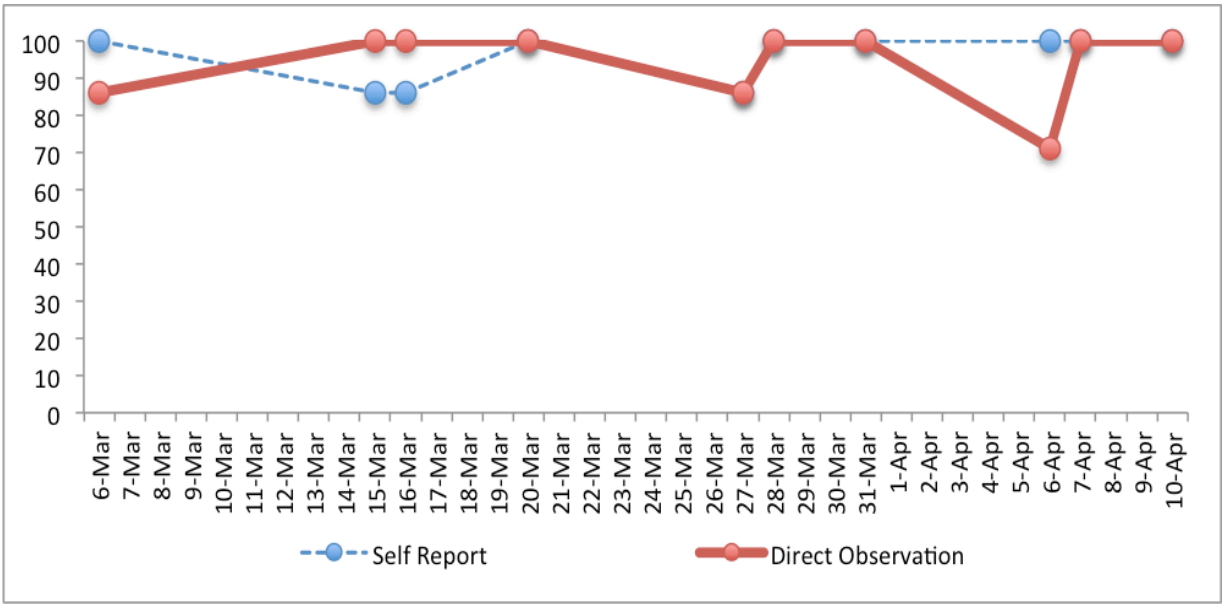


Figure 5. Self-report and direct observation ratings over the data collection period for Teacher 4.

Limitations and Future Directions

These results should be interpreted in the context of a number of limitations. First, as previously mentioned, the sampling plan (i.e., twice per week for five weeks) may not have been the most advantageous and future studies should have more frequent observations that occur closer in time to one another to limit any extraneous variables between observations that could affect implementation or teacher motivation. Another factor contributing to the low coefficients could be a combination of method and day effects. Because of the study design, it was impossible to tease apart the variance due to each method on its own as well as day of the week and the interaction between these two facets (i.e., Monday, Tuesday, etc.).

Additionally, throughout the study teachers were not given performance feedback on their implementation of the Response Card intervention. Aside from the initial training session and Response Card Protocol (Appendix C) that teachers were given during recruitment, no further feedback was given. Observers simply collected the self-report rating forms after the

intervention was implemented. Future studies would most likely benefit from a consultation piece, as this is best practice when working with teachers in schools. Consultation could be added into the model as a facet with three levels (e.g., no consultation, self-report and direct observation comparison group, and full feedback group) to examine which level of consultation coupled with which method of treatment integrity is most efficient at producing a reliable measure of treatment integrity the quickest.

Another factor affecting treatment integrity was teacher buy-in and motivation to accurately and attentively follow the intervention steps and complete the self-report form. To increase teacher buy-in, future studies utilizing the Response Card Intervention, or any intervention involving academic engagement, would likely benefit from a general classroom behavior assessment. With the Gresham et al. (2017) study, which utilized the Good Behavior Game intervention, positive results were made evident to teachers quickly as they could see that their students were responding to the intervention and exhibiting more prosocial behaviors as well as following classroom rules. These immediate positive effects likely motivated teachers to implement the steps of the intervention accurately and consistently. If teachers were made aware of the benefits of the Response Card Intervention, perhaps by having an objective measure of academic engagement rated during the intervention period, they would be motivated to accurately and consistently follow the steps of the intervention and report their performance.

Finally, the self-report and direct observations forms were, in hindsight, poorly constructed. These forms required teachers and observers to check "yes" or "no" to indicate whether the teacher performed each of the seven steps accurately. With the Response Card Intervention, steps 1 and 2 involve telling the class that the intervention is beginning and reviewing the rules of the intervention with the class. During the initial training, "reviewing the

rules of the intervention" was explained to teachers as well as outlined in the protocol as reviewing what the students should do when answering questions on response cards (e.g., answer independently, only write on white boards when instructed to do so, etc.). Some teachers reviewed more behavioral rules than procedural rules for step 2 (e.g., don't push down on markers, don't blurt out answers, don't draw on white boards, etc.). Although reviewing behavioral expectations for the intervention is necessary, step 2 caused some confusion for observers as to whether the step was accurately completed. Furthermore, steps 3 through 7 outline what the teacher should do when presenting one question from start to finish. Since the Response Card Intervention is meant to have teachers ask multiple questions during one implementation period, this also caused confusion for observers and teachers when rating whether the steps were completed. For example, a teacher could ask ten questions during the intervention and not follow the steps correctly for nine out of the ten questions, but the steps would still be rated as complete. Future studies should utilize more explicit directions and training for teachers and observers to avoid this confusion that likely contributed to the results of the present study. One method that could potentially solve this issue is instance rating, in which raters would be required to rate each "instance" of a behavior. In the context of a response card intervention, each question a teacher asked would be counted as one instance and would be rated accordingly.

In general, the mere construction of components checklists in regard to the scaling of measurement and setting of parameters is tricky and must be given proper attention or important details of the intervention will fall through the cracks. Without a solid foundation for which to measure treatment integrity, the generalizability of the measure is moot.

Conclusions

Although the results of this study were not significant, the results coupled with the study's design and limitations add to the research literature, which utilizes Generalizability Theory to assess the reliability of behavioral assessment in schools (e.g., Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007; Hintze & Matthews, 2004; Volpe & Briesch, 2014) as well as the new concept started by Gresham et al. (2017), which encourages researchers to add to the treatment integrity database. The results should be interpreted with caution and more research is needed involving the Response Card intervention as well as different behavioral and academic interventions to allow for more solid conclusions to be made. It also opens up the dialogue for the many issues surrounding components checklist construction as well as the scaling and measurement of treatment integrity measures, such as self-report and observer report checklists. This study is an important step forward in creating the treatment integrity database that will benefit practitioners, researchers, and policy-makers when they are faced with the decision of which treatment integrity measurement to use for various treatments for various issues, as well as how many assessments are needed in order to yield the most reliable results most efficiently.

References

Briesch, A.M., Chafouleas, S.M., Riley-Tillman, T.C. (2016). *Direct behavior rating: Linking assessment, communication, and intervention*. New York: The Guilford Press.

Briesch, A.M., Swaminathan, H., Welsh, M., & Chafouleas, S.M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*, 13-35.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review, 36(1),* 63–79.

Christle, C.A. & Schuster, J.W. (2003). The effects of using response cards on student participation, academic achievement, and on-task behavior during whole-class, math instruction. *Journal of Behavioral Education, 12(3),* 147-165.

Chronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons, Inc.

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(3-4),* 327-350.

Ferguson, T. D., Briesch, A. M., Volpe, R. J., & Daniels, B. (2012). The influence of observation length on the dependability of data. *School Psychology Quarterly, 27(4),* 187-197.

Fiske, K. E. (2008). Treatment integrity of school-based behavior analytic interventions: A review of the research. *Behavior Analysis in Practice*, *1(2),* 19.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review*.

Gresham, Frank M. (2009). Evolution of the treatment integrity concept: Current status and future directions. *School Psychology Review, 38(4),* 533-540.

Gresham, F.M., Dart, E., & Collins, T. (2017). Generalizability of multiple measures of treatment integrity: Comparisons among direct observations, permanent products, and self-report. *School Psychology Review, 46(1),* 108-121.

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

Lambert, M.C., Cartledge, G., Heward, W.L., & Lo ,Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Inteventions, 8(2),* 88-99.

Lane, K. L., Boccian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential—but often forgotten—component of school-based interventions. *Preventing School Failure, 48(3*), 36–43.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of schoolbased interventions with children in the Journal of Applied Behavior Analysis: 1991–2005. *Journal of Applied Behavior Analysis, 40,* 659–672.

Mushquash, C., & O'Connor, B. P. (2006). SPSS, SAS, and MATLAB programs for generalizability theory analyses. *Behavior Research Methods, 38*, 542-547.

Noell, G. H., Witt, J. C., Slider, N. J., & Connell, J. E. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review*, *34*(1), 87.

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. DOI 10.1007/s10864-013-9177-1

Perepletchikova, F., Treat, T., & Kazdin, A. (2007). Treatment integrity in psychotherapy research: Analysis of studies and examination of associated factors. *Journal of Consulting and Clinical Psychology, 75,* 829–841.

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.

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

Volpe, R. J., & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41(3),* 246–261.

Wickstrom, K., Jones, K., LaFleur, L., & Witt, J. (1998). An analysis of treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 13*, 141-154.

Appendix A: Direct Observation Form

Observer Report Form

1. The teacher states to the class that it is time to respond using response cards.	Y N
2. The teacher reviews the rules of the intervention with the class.	Y N
3. The teacher presents a question to the class.	Y N
4. The teacher provides adequate time for students to write their answers on the response cards.	Y N
5. The teacher requests that students present their cards.	Y N
6. If less than 75% of students provide correct answers, the teacher instructs the class to fix their answers and provide a rationale for the answer; if 75% or more of the class answers correctly, the teacher reveals the correct answer to the class.	Y N
7. The teacher provides verbal praise for correct responses.	Y N

Observer _____ Date _____

Teacher _____

Integrity = ____/____ = _____% IOA?

Appendix B: Teacher Self-Report Form

Teacher Self-Report Form

1. The teacher states to the class that it is time to respond using response cards.	Y N
2. The teacher reviews the rules of the intervention with the class.	Y N
3. The teacher presents a question to the class.	Y N
4. The teacher provides adequate time for students to write their answers on the response cards.	Y N
5. The teacher requests that students present their cards.	Y N
6. If less than 75% of students provide correct answers, the teacher instructs the class to fix their answers and provide a rationale for the answer; if 75% or more of the class answers correctly, the teacher reveals the correct answer to the class.	Y N
7. The teacher provides verbal praise for correct responses.	Y N

Teacher _____ Date _____

Integrity = ____/____ = _____%

Appendix C: Teacher Training Protocol

Response Card Intervention Protocol

Study begins: Monday, 3/6/17

Materials: You have been given enough "white boards," socks, and markers so that each child will receive one of each. These are yours to keep if you wish.

Intervention period: _____

Beginning March 6th, you will implement the Response Card Intervention in the instructional block that you chose (indicated above). You will conduct class as you normally would during this period. The only difference is that for whole group instruction, instead of the students responding verbally they will write their answers on their cards (see steps below). You will do this every day for 5 weeks, ending on Monday, April 10th. During these five weeks, two days of each week chosen at random will be observed by trained graduate students from LSU. When the graduate student observes they will fill out an observation form. They will hand you a self-report form, which is identical to the observation form they will be filling out, that you are to fill out while implementing the intervention. This self-report form lists the steps of the intervention. Whenever you complete a step, you will circle "Y" indicating that you completed the step. You will hand this completed form to the observer when they leave.

Steps of Intervention

1. Teacher states to the class that it is time to respond using response cards.
    a. Students will take out their white boards, sock (used as eraser), and marker.
2. The teacher reviews the rules of the intervention to the class.
    a. The students are to write down their answers on their white boards when indicated by the teacher. They are to think of their answers independently.
3. The teacher presents a question to the class.
    a. This does not have to be a formal process. These are questions that would come naturally with the lesson and that you would normally ask the class. Instead of choosing a particular student to answer or having the whole class respond at once, each student will write their answer on the response card.
4. The teacher provides adequate time for students to write answers on their response cards.
5. The teacher requests that students present their cards.
6. If less than 75% of students provide correct answers, the teacher instructs the class to fix their answers and provide a rationale for the answer; if 75% or more of the class answers correctly, the teacher reveals the correct answer to the class.
    a. This can be a quick scan of the classroom; it does not have to be exact. If you think that less than ¾ of the class answered correctly, you will have them try again. If the majority answered correctly, you will reveal the correct answer.
7. The teacher provides verbal praise for correct responses.

Appendix D: Graduate Research Assistant Protocol

Response Card Treatment Integrity G Theory Study
Direct Observation Protocol

Schedule: The study will run for five weeks (25 days). Ten observations of implementation of the Response Card Intervention must be completed for each teacher (6). Therefore, 60 observations must be conducted over the course of the project. Furthermore, 30% of those implementations will require interobserver agreement (IOA). That means a second observer will be present for 3 observations for each teacher. You will find the observation schedule in the dropbox folder. This schedule outlines the times that each teacher will be implementing the intervention, observation days, and observers for each time slot. IOA times are to be filled out by Thursdays at 10pm. This will leave time for any schedule changes on Friday so that the schedule can be finalized before the weekend. Each observation will last for 25 minutes. In the case that the teacher and students are absent from school (e.g., fieldtrips, holidays), the observation schedule will be pushed back the same number of days as the absence. We can adjust the end of the 5-week period as we come across absences.

Observation Procedure: To determine the observation time for the teacher you will observe, check the schedule that is located in the dropbox folder. If there are any complications, email the group by 10pm on the Thursday before.

Materials: Observer Report Form
Teacher Self-Report Form
Writing utensil

Arrive to the classroom with the necessary materials at least 5 minutes before the Response Card Intervention is scheduled to be implemented. It is critical to the study's success that the observer be present for the entire observation period in which the Response Card Intervention is scheduled. This will require adjusting your travel time for traffic and reasonably predictable events.

When you enter the classroom, do so with minimal interruption, hand the self-report form to the teacher, and position yourself in an area that will not cause any disruption throughout the entire observation period. You may sit or stand, whichever you prefer, as long as you are comfortable, have an unobstructed view of the classroom, and can easily record teacher behavior.

Your 25-minute observation begins at the time indicated on the observation schedule. Your first responsibility will be to look for components 1 and 2 on Observer Report Form, as these should be completed by the teacher before the intervention begins. Once these steps have been completed (or omitted) you should focus primarily on recording whether or not the teacher completes components 3, 4, 5, 6, and 7. Do not share this data with anyone, especially the teacher or other observers. Continue this process until the observation period ends. At this time, the teacher should have completed components 1 through 7 on the self-report form you handed

her before the intervention began. **DO NOT AT ANY TIME REMIND THE TEACHER TO COMPLETE ANY INTERVENTION STEPS.**

When the 25-minute observation period is over, even if the teacher has not completed all the steps, approach the teacher discreetly to collect the self-report form.

Do not leave the classroom without the self-report form. It is critical that this data is collected immediately after the end of each Response Card Intervention session that you observe.

Ensure you have written all observation information on the forms, including the date and time of the observation, your initials, and the teacher's name. This is especially important to note on the Teacher Self-Report Form because the teacher does not record this information.

ACTION ON EXEMPTION APPROVAL REQUEST

Institutional Review Board
Dr. Dennis Landin, Chair
130 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu | lsu.edu/irb

TO: Elizabeth Wilson
Psychology

FROM: Dennis Landin
Chair, Institutional Review Board

DATE: November 15, 2016

RE: **IRB#** E10240

TITLE: Generalizability of Multiple Measures of Treatment Integrity: An Empirical Replication

New Protocol/Modification/Continuation: New Protocol

Review Date: 11/15/2016

Approved X **Disapproved**

Approval Date: 11/15/2016 **Approval Expiration Date:** 11/14/2019

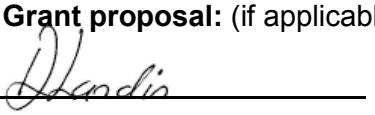
Exemption Category/Paragraph: 1; 2a,b

Signed Consent Waived?: No

Re-review frequency: (three years unless otherwise stated)

LSU Proposal Number (if applicable):

Protocol Matches Scope of Work in Grant proposal: (if applicable)

By: Dennis Landin, Chairman 

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants, including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.
8. **SPECIAL NOTE: When emailing more than one recipient, make sure you use bcc. Approvals will automatically be closed by the IRB on the expiration date unless the PI requests a continuation.**

* *All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb*

Vita

Elizabeth Kelsey Wilson was born in Baton Rouge, Louisiana and attended Louisiana State University in Baton Rouge from August 2008 to May 2012 where she earned a Bachelor of Arts in psychology. She then attended the University of Texas at Dallas in Richardson, Texas, from August 2012 to May 2014 where she earned a Master of Science in psychological sciences with a focus in developmental psychology from the School of Brain and Behavioral Sciences. Elizabeth is a third-year graduate student in school psychology at Louisiana State University under the supervision of Dr. Frank M. Gresham. Elizabeth's clinical and research interests include behavioral interventions for students with emotional and behavioral disorders, mental health awareness, as well as the influence of trauma and violence exposure on students.