

University of Nevada, Reno

Treatment Integrity Reporting in JABA: 1980-2019

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science in Psychology

by

Alisha L. Holder

Dr. Patrick M. Ghezzi, Thesis Advisor

May, 2021

Copyright by Alisha L. Holder 2021
All Rights Reserved



THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

ALISHA L. HOLDER

entitled

**Treatment Integrity Reporting in
JABA: 1980-2019**

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Patrick Ghezzi, Ph.D.
Advisor

Matthew Lewon, Ph.D.
Committee Member

MaryAnn Demchak, Ph.D.
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

May 2021

Abstract

The purpose of the present investigation was to extend the Gresham et al. (1993) study on the frequency of reporting data on treatment integrity (TI) in research involving children published in the *Journal of Applied Behavior Analysis*. The Gresham et al. study began in 1980 and ended in 1990; the present began in 1991 and ended in 2019. The results indicate that the percentage of articles with TI data has increased over the years from just 16% in 1980-1990 overall to roughly 50% today. Two additional variables were included in the analysis, the setting or location in which the research was conducted, and the person responsible for implementing treatment. Schools were identified as the setting in which TI data were most frequently obtained, and researchers were identified as the people from whom TI data were most frequently obtained. The implications of the results of the present study are discussed in the context of the significance of collecting, analyzing, and reporting TI data in the field of applied behavior analysis.

Acknowledgements

To my advisor, Dr. Patrick M. Ghezzi, thank you for your never ending guidance and support throughout my graduate career, and for sticking with me and this project and seeing it through to the very end. My knowledge base of not only “knowing about” but “knowing how” is truly dedicated to your teachings and will forever be the backbone of my future career.

I would also like to thank Isha Patel, who contributed to the initiation of this extremely important topic and to Vanessa Willmoth for her collaboration in the development of this research endeavor. I am honored to say the completion of this tremendously crucial research has been one of the most rewarding aspects of my graduate career. I hope the findings can help promote changes that are capable of steering the field in the right direction as it continues to grow and evolve.

To my husband, Michael J. Churchfield, this entire process would not have been possible without your true selflessness and encouragement each and every day. I will forever be grateful and cannot say enough about how instrumental you were in helping me achieve my goals.

And lastly, to my parents, who allowed me to pursue my dreams and craft my knowledge, leading to a future career that I love, all the while being across the country, I cannot thank you enough. You stood by patiently through each and every set back and achievement and I can finally say “I did it.”

Table of Contents

Abstract.....	i
Acknowledgements.....	ii
List of Figures.....	iv
Introduction.....	1
Previous Research.....	4
Method.....	8
Phase I.....	8
Phase II.....	9
IOA.....	10
Results and Discussion.....	11
General Discussion.....	14
Appendix A.....	26
<i>Checklist/ Data collection tool I (1991-2019)</i>	26
<i>Checklist/ Data collection tool II (2011-2019)</i>	26
Appendix B.....	27
<i>Reliability Table: Treatment Integrity</i>	27
Appendix C.....	28
<i>Reliability Tables: Location and Implementation Data</i>	28
Tables and Figures.....	29
<i>Table 1. Treatment Integrity Data: (1991-2019)</i>	29
<i>Figure 1</i>	30
<i>Figure 2</i>	31
<i>Figure 3</i>	32
<i>Figure 4</i>	33
<i>Figure 5</i>	34
<i>Figure 6</i>	35
<i>Figure 7</i>	36

List of Figures

- Figure 1.* Continuation graph presented with a phase line between the Gresham data and the most recent data.....30
- Figure 2.* A visual representation of treatment integrity reporting from 1991-2010 including trend lines.....31
- Figure 3.* A closer look at integrity data from the past 9 years from 2011-2019 including most recent trend directions and slopes.....32
- Figure 4.* Chart showing the percentage of articles including integrity data across the settings/locations where the studies were conducted each year.....33
- Figure 5.* Chart showing the percentage of articles including integrity data across the settings/locations where the studies were conducted as a combined representation including all articles published in JABA from 2011-2019.....34
- Figure 6.* Chart showing the percentage of articles reporting integrity data according to who implemented the procedures each year.....35
- Figure 7.* Chart showing the percentage of articles including integrity data reporting according to who implemented the procedures.....36

Introduction

Accurate observations and precise measures of the dependent variable are defining features of experimental research in the natural sciences. Research in the applied science of behavior, Applied Behavior Analysis (ABA), reflects this truism by reporting inter-observer agreement (IOA) with respect to the occurrence and nonoccurrence of a given “target” behavior (Hartman, 1977; Mudford, Taylor, & Martin, 2009; Volmer, Sloman, & Pipkin, 2008). It is a cardinal aspiration in ABA, in fact, to describe the dependent variable so accurately that any two independent observers would always agree on the responses they saw and did not see, i.e., 100% IOA.

The practice of collecting IOA on the dependent variable guards against the potential for bias and error when humans as opposed to machines serve as the “transducers” for observing and recording behavior (Johnston & Pennypacker, 2009). With IOA, a researcher can conclude with greater (or lesser) confidence that a change in a target behavior is a function of the relevant independent variable(s) as opposed to the mistakes an observer made in applying a behavior observation code (Page & Iwata, 1986). IOA serves this purpose by means of determining the agreement between independent observers with respect to the correct occurrence and nonoccurrence of a given stimulus at a given place and time.

The independent variable is subject to a similar analysis. Known in ABA as “treatment integrity” or simply TI, the area exposes the problems inherent in instructing humans to present, withhold, and remove stimuli as opposed to programming these events electronically. The extent to which a person actually follows a procedure or

protocol in a study involving motivational, antecedent, and consequential events is the top question, and human error and bias can cloud the answer. A research assistant in a study on motor response prompting, for example, could misunderstand the procedure regarding when, where, and at what level of assistance to prompt the target response. It would be important for the researcher to know this, and better yet, to have an objective measure of the extent to which the written description of the procedure matches the behavior of the assistant in giving the right prompts at the right times. Conclusions regarding the functional relationships between an independent variable and a target response are stronger with high IOA, and generalizations based on this strength are bound to be more credible than research with little or no regard for an objective measure of TI.

In reference to the implementation of an intervention protocol, the terms procedural integrity and treatment integrity are often times used interchangeably; however, a distinction is often made between the two. Treatment integrity is in fact a type of procedural integrity, but it exclusively refers to “the extent to which a treatment is implemented as planned,” (Yeaton & Sechrest, 1981). On the other hand, the term procedural integrity can be used in reference to any research procedures or protocols and is not limited to the experimental phase of a study. The focus of the present study is treatment integrity and is aimed at evaluating the extent to which researchers report data related to the implementation of the independent variable/s.

Systematic replication, the backbone of science, is most meaningful when a current researcher can count on past researchers not only to provide detailed descriptions of their treatment protocols but also to measure the extent to which the protocols were implemented as described (Sidman, 1960; Pennypacker, 1980). Locey (2020) addressed

the “replication crises in psychology” and raised the possibility that ABA, despite the common practice of replicating procedures within the same study, is still vulnerable to replication failures between studies. According to Locey in 2020, as more and more applied studies with large effects are published each year, the risk for procedural replication errors between studies increases dramatically¹. Collecting and reporting TI data on a regular basis reduces this risk, in theory, by the consistent results produced by faithfully replicating the procedures from previous studies, which themselves report a high degree of fidelity with respect to the independent variable.

As well, a growing body of evidence indicates that ABA procedures implemented with a high TI produce better outcomes compared to procedures with lower TI (Fiske, 2008; Fryling, Wallace, & Yassine, 2012; McIntyre et al., 2007; Sanetti & Kratochwill, 2009; Wilder, Atwell, & Wine, 2006). This alone is a compelling reason for ABA researchers to collect, analyze, and report TI data routinely, and yet the practice is not as prevalent as might be expected.

This conclusion is based on the research literature to date on TI. It is a controversial literature, one that harbors a critical view of ABA with respect to the slow progress over the years in increasing the frequency of TI data reported in research articles appearing in major journals such as the *Journal of Applied Behavior Analysis (JABA)*. We examine this literature next.

¹ The mean number of studies published in *JABA* with a focus on children and youth grew from 14 per year in 1980-90 to 47 per year in 2010-19.

Previous Research

Billingsley and his colleagues (1980) were among the first to investigate the extent to which TI is reported in ABA journals. They examined 108 research articles published in *JABA* between 1977 and 1978 and in *Behavior Modification* between 1978 and 1979. Between the two journals, just 5.6% of the articles included a measure of TI compared to the 82% that included a measure of the dependent variable. The greater emphasis given to ensuring the integrity of the dependent variable alarmed Billingsley et al., because it raised the possibility that conclusions, generalizations, and recommendations about the effects of a given treatment or intervention in ABA may be mistaken and even potentially harmful to consumers.

Peterson and her colleagues (1982) directed their analysis to TI data reported in research articles published in *JABA* between 1968 and 1980. Of the 539 research articles included in their analysis, roughly 20% provided a measure of TI. “The majority of articles published (in *JABA*),” concluded Petersen et al., “do not use any assessment of the actual occurrence of independent variables and a sizable minority do not provide operational definitions of the independent variable.” This neglect, wrote Peterson and her colleagues, “has no place in a science of behavior.”

Imploring their fellow researchers and journal editors to establish better TI practices in the future, Peterson et al., (1982) offered some suggestions about how to navigate the barriers that deter researchers from collecting the data. One suggestion involved assigning a single observer to collect IOA on both the dependent and independent variable(s). This is common practice today among ABA researchers, that is, to collect both types of data at the same time and analyze the data later for IOA purposes.

A recent example is a study by Bachmeyer et al., (2019) on improving the eating behavior of three young children with disabilities and a feeding disorder. All experimental sessions were videotaped and later analyzed by trained, independent observers with respect both to the targeted behaviors (eating, swallowing) and the procedures (reinforcement, extinction) used to change the behaviors.

Gresham and his colleagues (1993) examined *JABA* with respect to research involving children and youth published over a 10 year time period, spanning from 1980 to 1990. The authors discovered that just 16% of the 158 articles included a measure of TI and found that two thirds failed to define the independent variable with enough detail to permit replication. The remaining one third of the studies did provide detailed information on the independent variable, which Gresham et al. cited as progress relative to the results reported 10 years earlier by Petersen et al., (1982). Still, Gresham and his colleagues concluded that, “Peterson et al.’s call for increased measurement of independent variables has not been heeded.”

More recently, McIntyre, Gresham, DiGennaro, & Reed (2007) examined 142 articles, including 153 individual studies published in *JABA* between 1995 and 2005 for evidence of TI data in research involving school-based interventions for children and youth. While less than a third (30%) of the articles provided TI data, most (95%) provided a reasonably clear definition of the independent variable. This suggests significant improvement, on one hand, in defining the independent variable in such a manner that would lend itself to reporting TI data, but little improvement, on the other hand, in actually reporting the data.

McIntyre et al., (2007) took their analysis a step further by targeting the personnel involved in implementing procedures to determine whether they had a role in reporting TI data, and to determine if TI data were more or less likely to be reported as a factor of who was responsible for protocol implementation. They found that classroom teachers, as opposed to researchers and research assistants, were most involved in the process of collecting TI data. These are important data because they identify the places and people involved in a study who are more or less likely to participate in the TI process, and in turn determine the people who might benefit the most from further training, and supervision on the topic.

One last study in this area, by Armstrong and her colleagues (1997) is a younger version of the study by Billingsley et al., (1980). The Billingsley study revealed a shortcoming in reporting TI data in *JABA* from 1977-1978 and found the same shortcoming in the journal *Behavior Modification* from 1978-1979. Armstrong et al. analyzed 39 ABA studies published from 1991-1994 in the *Journal of Developmental and Physical Disabilities*. Of these, a paltry 23% contained TI data. The figure is high, actually. To the studies that reported TI data, Armstrong et al. added studies that simply “assured” the reader that the intervention was implemented as described. Mere assurances, in our view, are no substitute for real TI data.

A recent development in education pertains to the guidelines for conducting and evaluating applied research in education. The What Works Clearinghouse (WWC), established in 2002 by The Institute for Education Sciences (Kratochwill et al., 2013) for large-N studies, now includes evaluative criteria for single-case research. TI data reporting is not among the criteria, an omission that Wolery (2013) cast as a request to

take a “leap of faith” each time a study is published. The leadership of WWC agreed yet placed the responsibility on the researcher and not WWC to evaluate TI. As Hitchcock et al., (2015) put it, “primary researchers and not WWC coders are in a much better position to examine why a promising intervention may not have yielded desirable outcomes, and if fidelity concerns is suspected, what to do next” (pg.148). While the WWC does not directly include this measure among their standards, other evaluative criteria used in the field of special education does. For example, Horner and his colleagues in 2005 published guidelines for identifying various defining features of evidence based practices in single-case research. Among the quality indicators provided, the authors explicitly state that “overt measurement of the fidelity of implementation for the independent variable is highly desirable” and is in fact “expected either through continuous direct measurement of the independent variable or an equivalent.” (Horner, Carr, Halle, McGee, Odom, & Wolery, 2005). Within the behavioral literature, Tate and his colleagues (2016) put the matter in the hands of ABA journal editors and recommended that TI data be required as a condition for publication.

The present study extends the line of research began by Gresham and his colleagues (1993) on TI data reported in *JABA* involving children and youth. The present analysis starts with articles published in 1991, the year Gresham et al., (1993) ended their analysis, and proceeds through 2019. Our first aim was to update the field on what has transpired over nearly three decades of research published in *JABA*. Our second aim was to follow McIntyre and his colleague’s (2007) lead by identifying the personnel involved in implementing protocols and determining the role they may play in reporting TI data.

Method

Research articles published in *JABA* between 1991 (Vol. 24) and 2019 (Vol. 52) constituted the subject of the present analysis. The articles involved research with children and youth between birth and 19 years only. A grand total of 1,077 articles met the inclusionary criteria (see Appendix A).

There were two phases in the study. Phase I, was a direct replication of the methods used by Gresham et al., (1993) for both the inclusion of articles and the examination of TI data. Phase II built on the study by McIntyre et al., (2007) on identifying the person(s) responsible for measuring and reporting TI data published from 1995-2005 in *JABA*. This second phase involved research published in *JABA* involving children and youth between 2011 (Vol. 44) and 2019 (Vol. 52). A total of 200 articles, all coded as “yes” in phase I were identified for inclusion in this phase of the present study.

Phase I

Following the procedure outlined by Gresham et al., (1993), an article was entered into a spreadsheet according to the participant(s) age, the year of publication, and TI data. Age was recorded as it was reported in the article, as was the year of publication. TI data were coded and sorted into one of three mutually exclusive categories: (1) Yes, (2) No, and (3) Monitored. (For information regarding the checklist used to code articles for TI, see Appendix A).

“Yes” was coded if the article provided detailed information on treatment protocols and reported an objective value such as IOA with respect to a written task analysis of a procedure and the implementation of each component of the procedure (e.g., Lechago, Carr, Grow, Love, & Almason, 2010).

“No” was coded if the article did not include information on TI and did not provide objective values relevant to it (e.g., Hanney & Tiger, 2012).

“Monitored” was coded if an article “kept track of” TI. Included in this category were articles that (1) excluded values relevant to TI, (2) instructed the reader to look elsewhere for the TI data (e.g., contact the first author), or (3) reported TI data for one experiment but not for the other experiment(s) in articles involving multiple experiments. (e.g., Normand, Machado, Hustyi, & Morley, 2011).

Phase II

This phase of this analysis centered on articles in *JABA* published between 2011 and 2019 and scored “Yes” from Phase 1. The two variables of interest included (A) the location of the study and (B) the identity of the person(s) responsible for implementing a procedure or protocol. The location data were coded and sorted into one of five settings: (1) school, (2) clinical, (3) home, (4) university, and (5) other (e.g., community). If an article included more than one setting, each setting was recorded. (For information regarding the checklist used to code the location data, see Appendix A.)

The same articles were coded a second time on the basis on who implemented the protocol or procedure in the study. “Implementers” were coded and sorted into one of six categories: (1) experimenter/researcher, (2) investigator/author, (3) teacher/instructor, (4) parent/caregiver, (5) therapist/clinician, and (6) other. If multiple implementers were listed in the article (e.g., therapist and caregiver), each one was included in the analysis. It was possible, therefore, to obtain a greater number of implementers than articles for any one year. If it was clear that the “implementer” served dual roles (e.g., classroom teacher, first author), the article was coded and sorted according to the role they held at

the time they implemented a protocol or procedure. A classroom teacher, for example, may be an author on a publication but implemented a procedure as the teacher. In this case, the implementer was coded and sorted as a teacher, not an author. (For information regarding the checklist used to code the implementer data, see Appendix A.)

IOA

The primary reviewer (the first author) coded and sorted 100% of the research articles in the present study. An independent reviewer designated as the secondary reviewer, trained by the primary reviewer on the Phase I and Phase II procedures, randomly selected no less than 20% of articles each year for a second round of coding and sorting. IOA measures were obtained for each year for both phases of the study by dividing the total number of agreements between the primary and secondary reviews with respect to the Data Checklist for Phase I and II by the total number of agreements plus disagreements multiplied by 100%.

Appendix B displays the Phase 1 IOA data for each year from 1991 through 2019. The percentage of articles on which IOA was calculated each year ranged from 20% to 23.5%, and the IOA values averaged 95.6% over the entire 29 years; for 20 of those years, IOA was 100%. The year 2005 was an outlier, where IOA fell to 71.4%.

Appendix C shows the Phase II IOA data for each year from 2011 through 2019. The results of the total IOA for location/settings was 92.5% over this time period and ranged from 80%-100% each year. The total number of locations reported for at least 20% of articles coded as “yes” were compared. Matches included all locations that were recorded by both observers for the same article and non-matches included both instances where observers recorded different locations in addition to instances where one observer

listed a location that the other observer omitted. Both types were scored as a non-match. Since multiple locations could have been listed for any given article, reliability was calculated dividing the total number of location agreements divided by the total number of agreements plus disagreements between observers and multiplied by 100%. The total IOA results for implementers was 91.4% over this same time period and ranged from 71.4%-100% each year. The total number of implementers reported for at least 20% of articles coded as “yes” were compared. The method for scoring matches and non-matches was identical to the way in which the location IOA data was obtained and calculated. Details on what constituted a match vs a non-match is described above.

Results and Discussion

Table 1 contains the raw numbers and percentages of articles for each year (1991-2019) and for each one of the three TI categories, Yes, No, and Monitored. Of the 1,077 studies included in the first phase of analysis, 36.3% (N=391) were coded as Yes, 56.3% (N=606) were coded as No, and 7.4% (N=80) were coded as Monitored.

Figure 1 shows the distribution of the three TI categories over time in graphic form. Note the phase change line in Figure 1. The line marks, for comparative purposes, the end of the 1980-1990 study by Gresham et al., (1993) and the beginning of the present study, from 1991 through 2019. (N.B. The Gresham et al. data were taken from Table 1 of their 1993 publication.)

One notable feature of the data in Figure 1 is the abrupt rise in 1991 in the percentage of articles that reported an objective value of TI. Since Gresham et al. published their data two years later, in 1993, it is difficult to credit them for the 46% increase in TI reporting from 1990 to 1991. It is conceivable that the data appeared at a

conference or convention prior to publication, that members of the editorial board of *JABA* were suddenly alerted to the importance of reporting TI data, and that changes ensued in the journal's review and editorial process that favored studies that reported TI data when evaluating research for publication. It is clear, in any case, that the post-1990 abrupt increase was only temporary. The percentage of articles with TI data declined after 1991, and by 2000, the percentages had again reached levels within the 10-20% range reported by Gresham et al., (1993) for the entire decade.

A second feature in Figure 1 is the steady increase in TI reporting that occurred following the low rates observed in the year 2000. With the exception of 2006, rates never fell below 20%, and in fact, they continued to climb through 2011 to just over 40%. Following one small, short-lived decline in 2012 and 2013, TI reporting climbed to over 60% by 2019, the last year of the present study and the highest value obtained in 29 years.

A third feature in Figure 1 pertains to the percentage of articles reporting TI as "monitored." On average, under 10% of both the studies examined by Gresham et al., (1993) and the studies in the present analysis met the criteria. At approximately 3% for the 10 most recent consecutive years (2010-2019), the practice of "monitoring for TI" is as uncommon today as it has been for the past 29 years.

Figure 2 provides a closer look at trends in TI reporting from 1991-2019 providing a clear picture of TI data reporting in *JABA* immediately following the years in which the Gresham et al. data concluded in 1990. The promising news is that the practice of TI reporting has trended upwards since 1991. However, the practice of reporting no TI data is also trending upwards, with an ever so slight positive slope. The most concerning

feature of this data is that “No” reports of TI data still outnumber “Yes” reports by roughly 15 percentage points.

Figure 3 is a close-up view of recent trends in TI reporting from 2011 through 2019. The results are quite different from previous evaluations of past periods, and a shift in the direction toward greater TI reporting finally emerges. The positive slope in the trend toward including TI data is comparatively steeper than previously observed, and the trend for excluding TI data is trending downward direction, giving both reporting TI data and not reporting TI data similar slopes but in diverging directions.

The results of Phase 1 point to a substantial increase over time in the number of studies published in *JABA* involving children and youth that include TI data. Gresham et al., (1993), by comparison, found that 16% of the studies in *JABA* included TI data from 1980-1990, while 48% of the studies from 2011-2019 included TI data, a 32% increase between the two time periods.

The results of Phase II appear in Figures 4-7. The data are a 2011-2019 subset of the 1991-2019 data obtained from the pool of studies that reported TI data in Phase 1. Bear in mind that Phase II also consists only of studies involving children and youth.

Figure 4 shows the various locations of research reporting TI data each year from 2011-2019. The highest percentage of locations is school, which accounted for 49% of the TI data reported from all locations. In every year, the number of studies including TI data were most likely to be conducted in school settings with the exception of 2016 in which clinical settings outnumbered school locations. Of the articles reporting integrity measures, home, university, or other settings were found to account for the least

percentage of articles for each year. Figure 5 represents the total percentage of articles reporting TI data in each setting over the entire time frame from 2011-2019.

Figure 5 compares the percentage of research locations over the same period. School predominated at 49%, followed by clinics at 30% and home-based locations at 15%. University locations and other locations amounted to 4% or less of all studies including TI data from 2011-2019.

Figure 6 shows the people responsible for conducting an experimental protocol or procedure (the “Implementers”) from 2011-2019. Given the predominance of schools as the top location for obtaining TI data (see Figures 4 and 5), it follows that studies involving teachers might be more likely to obtain TI data. Instead, of the total number of articles reporting TI data each year, experimenters/researchers constituted more than doubles the percentage of teachers for these articles, and in the majority of years, clinical staff comprised a higher percentage of the articles reporting TI data each year than teachers as well.

Figure 7 compares the percentage of implementers for all studies combined over the 2011-2019 time period. Researchers and experimenters accounted for a total of 51% of the TI reporting followed by therapists and clinical staff at 20% and teachers and instructors at 12%.

General Discussion

Gresham and his colleagues (1993) were the first to sound the alarm on the low incidence of TI reporting in research published in *JABA* involving children and youth. The results from their study showed that less than 20% of the articles provided TI data from 1980-1990. The present study replicated and extended the analysis from 1991-2019.

The results showed that 36% of the articles published in *JABA* over this period reported TI data. It is a modest increase, yet a closer look at the trends in the data from 2000-2019 and 2011-2019 especially shows an encouraging upswing to over 60% of articles with TI data in 2019.

The favorable comparison between reporting and not reporting TI data notwithstanding, the fact remains that the level of reporting TI data is still well below the level of reporting IOA data on a target behavior(s), i.e., the dependent variable. Those levels have been reported by Mudford, Taylor, & Martin in 2009 to be at or near 100% for continuously recorded behavior. Their study evaluated studies published from 1995-2005, however, a quick search of the 45 research articles included in the present study and published in 2019 showed that 100% of the studies reported objective measures on the dependent variable. Indeed, entire sections of most articles gave detailed descriptions of the target response(s), the procedure for obtaining and measuring IOA with respect to the response(s), and the results of various calculations (means, ranges), yet give relatively little or no space to TI data (e.g., Briggs, Lessor, Kamana, & Jess, 2019). The divide between reporting TI data on the independent variable and reporting IOA on the dependent variable remains wide and closing the gap between the two is a major goal for the future.

One way to achieve this goal may be to change the way the literature refers to the measurement and monitoring of these variables. For example, IOA typically refers to the agreement between two or more observers collecting data on dependent variables and is used to ensure reliability of these data. TI is reserved for evaluating the implementation of a treatment or intervention. The latter does provide more information to researchers

about the implementation of treatment protocols; however, the reliability of these data are lacking if IOA is not applied. Applying the same measurement and observation techniques to well defined components of both dependent and independent variables and ensuring the reliability of these measurements through IOA calculations applied to these measures can close the gap. Until the language with respect to these variables becomes more accurate and consistent, equal attention to both variables may continue to be inhibited.

Sidman (1960) implied long ago that the future of behavior analysis rests with establishing the generality of an experimental outcome. In ABA, the route to establishing generality begins and ends with systematic replication. The route to systematic replication, in turn, is paved with the dual dimensions of behavior and behavior change. When confidence is high with respect to the occurrence of a given response in relation to a given stimulus designed to alter its occurrence, i.e., a successful replication, it strengthens the generality of an experimental outcome in two ways. It affirms the generality of a previous outcome, and it breeds future replications with variation on a present outcome, *ad infinitum*.

In order for a present researcher to replicate the conditions of a previous investigation with variation, the researcher must know enough about those past conditions to replicate with confidence. The results from the current study suggest that it may be difficult to rely on past research to supply this information in sufficient detail to enable replication with respect to the means by which behavior changes, i.e., the effects of an independent variable. As Worley (2013) put it, researchers are asked to take a “leap

of faith” each time a study is published without TI data and simply assume that the independent variable was consistently implemented as designed.

The increasing trend in reporting TI data found here raises the question as to whether the trend is limited to articles published in *JABA* or limited to research involving children and youth published in *JABA*. Future research, fashioned after the studies by Billingsley et al., (1980) and Armstrong et al., (1997), will answer these questions, but in the meantime, an unpublished master’s thesis by Kathryn Sharp (2020) at Saint Louis University suggests that trend is not limited either to *JABA* or to research involving children and youth.

The impetus for Sharp’s study was an article by Tate et al., (2016) in which they presented guidelines to assist ABA researchers in identifying the most important elements in a publishable study, TI data included. Sharp’s analysis focused on ABA studies published in *JABA* and a variety of other journals (e.g., *The Analysis of Verbal Behavior*, *Behavior Analysis in Practice*, etc.) from 2010-2019, and included adults as well as children and youth. Of the 281 articles in the pool, 48% reported TI data. Sharpe’s figure compares favorably to the 46% found over the same period in the present study and implies that the upward trend found here is widespread.

The results of the present study further show that 50% of all the studies that reported TI data from 2011-2019 were conducted in a school setting. The current analysis involved only children and youth, which makes school a most likely setting for research, and yet just 12% of these studies included a teacher or instructor as the “implementer” of the independent variable.

Instead, researchers and experimenters assumed this role most of the time in schools as well as in the clinics, homes, and other settings included in the present analysis.

This finding raises questions regarding the social validity of the treatments and interventions that take place in schools and classrooms. Generally, the most likely consumer of research conducted in classrooms would involve teachers or other school personnel. This lack of consideration for social validity, when evaluating effective solutions that are implemented in school based settings, limits the generality of the findings. Future researchers should take note of this and adjust their practices to involve personnel specific to the environment in all aspects of research conducted in all settings. In addition to including the fidelity of the independent variable implementation in single case research, Horner et al., (2005) also addresses the importance of social validity. The authors explain that social validity goals are further enhanced when studies are able to not only establish that the procedures “can be applied by typical intervention agents,” but that these implementers also report that intervention protocols are acceptable, feasible, effective, and would be continued after the support or expectations of researchers are removed.

The finding that experimenters and researchers, as opposed to classroom teachers and instructors, were responsible for implementing the independent variable is inconsistent with previous research focused on school-based interventions (e.g., McIntyre et al., 2007). In their study on school-based intervention in *JABA* from 1995-2005, McIntyre et al., (2007) reported data that showed of all studies that included treatment integrity data (n=46), 14(30%) were conducted by teachers, followed extremely close behind by researchers which accounted for only one less article, with 13(28%) of the

same 46 studies reporting TI data. Of the 200 articles including TI data in phase II of the current analysis from 2011-2019, a much larger percentage (51%) of articles that reported TI involved researchers, whereas teachers only accounted for 12% of the same studies. The discrepancy might owe to the fact their focus was on children in school-based interventions exclusively, which makes a teacher either a most likely implementer or most likely to be placed or sorted in that role as a function of simply being the classroom teacher. In any case, TI reporting is most likely to occur in a school setting and least likely to occur in other settings such as clinics and homes.

Regular and special education research dominates the literature with both resources and with the application of these resources when identifying and evaluating single case research on the basis of quality indicators. One example mentioned earlier, The What Works Clearinghouse, released a new Standards Handbook (Version 4.1) in 2020. It is interesting to find that it continues to include requirements related to outcome measurements (i.e., dependent variables) in all phases of a study but with respect to the independent variable, it merely requires that it be “systematically manipulated,” with the researcher in control of when and how conditions change. It is conceivable that the addition of TI reporting to the manifest of quality research indicators would elevate the school setting figures even further and closer to 100%. It is also conceivable that “spillover effects” could occur to ABA research conducted in other settings and with people of all ages. Editorial policies that outline the requirements for TI reporting in high profile journals such as *JABA* would elevate the figures quickly and, in doing so, would eliminate the problems and potential problems created by a loose regard for the fidelity of treatment in ABA.

References

- American Institutes for Research (AIR), & What Works Clearinghouse (ED).
(2020). What works clearinghouse™ standards handbook, version 4.1. What Works Clearinghouse.
- Arkoosh, M.K., Derby, K.M., Wacker, D.P., Berg. W., McLaughlin, T.F., & Barretto, A.
(2007). A descriptive evaluation of long-term treatment integrity. *Behavior Modification*, 31(6), 880-895.
- Armstrong, K. J., Ehrhardt, K. E., Cool, R. T., & Poling, A. (1997). Social validity and treatment integrity data: Reporting in articles published in Journal of Developmental and Physical Disabilities, 1991–1995. *Journal of Developmental and Physical Disabilities*, 9, 359-367.
- Bachmeyer, M. H., Kirkwood, C. A., Criscito, A. B., Mauzy, C. R., & Berth, D. P.
(2019). A comparison of functional analysis methods of inappropriate mealtime behavior. *Journal of Applied Behavior Analysis*, 52(3), 603-621.
- Billingsley, F., White, O. R., & Munson, R. Procedural reliability: A rationale and an example. *Behavioral Assessment*, 1980, 2, 229-241.
- Briggs, A. M., Dozier, C. L., Lessor, A. N., Kamana, B. U., & Jess, R. L. (2019). Further investigation of differential reinforcement of alternative behavior without extinction for escape-maintained destructive behavior. *Journal of Applied Behavior Analysis*, 52(4), 956-973.

- Florence D. DiGennaro Reed, & Coddling, R.S. (2014). Introduction: Advancements in procedural fidelity assessment and intervention: Introduction to the special issue. *Journal of Behavioral Education*, 23(1). 1-18.
- Fiske, K. E. (2008). Treatment integrity of school-based behavior analytic interventions: A review of the research. *Behavior Analysis on Practice*, 1, 19-25.
- Fryling, M. J., Wallace, M. D., Yassine, J. N., & Lerman, D. (2012). Impact of treatment integrity on intervention effectiveness. *Journal of Applied Behavior Analysis*, 45(2), 449-453.
- Gresham, F.M., Gansle, K.A., & Noell, G.H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis*, 26, 257-263.
- Hanney, N. M., & Tiger, J. H. (2012). Teaching coin discrimination to children with visual impairments. *Journal of applied behavior analysis*, 45(1), 167-172.
- Hartmann, D. P. (1977). considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 10(1), 103-116.
- Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, 11(6), 838-854.

- Hitchcock, J. H., Kratochwill, T. R., & Chezan, L. C. (2015). What works clearinghouse standards and generalization of single-case design evidence. *Journal of Behavioral Education, 24*(4), 459-469.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165-179.
- Ibañez, V. F., Piazza, C. C., & Peterson, K. M. (2019). A translational evaluation of renewal of inappropriate mealtime behavior. *Journal of Applied Behavior Analysis, 52*(4), 1005-1020.
- Johnston, J., & Pennypacker, H. (1980). *Strategies and tactics of human behavioral research*. Hillsdale, NJ: Erlbaum.
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26-38.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., Shadish, W. R., & What Works Clearinghouse (ED). (2010). Single-case designs technical documentation. *What Works Clearinghouse*.
- Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential—but often forgotten—component of school-based interventions. *Preventing School Failure: Alternative education for children and youth, 48*, 36-43.

- Lechago, S. A., Carr, J. E., Grow, L. L., Love, J. R., & Almason, S. M. (2010). demands for information generalize across establishing operations. *Journal of Applied Behavior Analysis*, 43(3), 381-395.
- Locey, M. L. (2020). The evolution of behavior analysis: Toward a replication crisis? *Perspectives on Behavior Science*, 43 (4), 655-675.
- McIntyre, L. L., Gresham, F. M., DiGenarro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the Journal of Applied Behavior Analysis 1991-2005. *Journal of Applied Behavior Analysis*, 40(4), 659-672.
- Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the Journal of Applied Behavior Analysis (1995–2005). *Journal of Applied Behavior Analysis*, 42(1), 165-169.
- Normand, M. P., Machado, M. A., Hustyi, K. M., & Morley, A. J. (2011). infant sign training and functional analysis. *Journal of Applied Behavior Analysis*, 44(2), 305-314.
- Northup, J., Fisher, W., Kahang, S. W., Harrell, R., & Kurtz, P. (1997). An assessment of the necessary strength of behavioral treatments for severe behavior problems. *Journal of Developmental and Physical Disabilities*, 9(1), 1-16.
- Page, T. J., & Iwata, B. A. (1986). Interobserver agreement: History, theory, and current methods. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (99– 126). New York: Plenum.

- Patel, I. K. (2014). Treatment fidelity in the journal of applied behavior analysis: 1991-2010. *Unpublished honor's thesis*, The University of Nevada, Reno, NV.
- Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis*, 15(4), 477-492.
- Plavnick, J. B., Ferreri, S. J., & Maupin, A. N. (2010). The effects of self-monitoring on the procedural integrity of a behavioral intervention for young children with developmental disabilities. *Journal of Applied Behavior Analysis*, 43(2), 315-320.
- Sanetti, L. M. H., & Kratochwill, T. R., (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School of Psychology Review*, 38(4), 445.
- Sharp, K. (2020). Behavior analytic research reporting strategies: An examination into ABAI journal publications from 2010 to 2019. *Unpublished master's thesis*. Saint Louis University, St. Louis, MO.
- Sidman, M., 1923. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.
- Skinner, B. F. (1953). *Science and human behavior*. New York: The Free Press.
- Tate, R.L., Perdices, M., Resenkoetter, U., McDonald., Togher, L., Shadish, W., Vohra, S. (2016) The single-case reporting guideline in behavioral interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology*, 4(1), 10-31.

- Vollmer, T. R., Sloman, K. N., & St Peter Pipkin, C. (2008). Practical implications of data reliability and treatment integrity monitoring. *Behavior Analysis in Practice*, 1(2), 4-11.
- Wilder, D. A., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis*, 39, 369–373.
- Wodarski, J. S., Feldman, R. A., & Pedi, S. J. (1974). Objective measurement of the independent variable: A neglected methodological aspect in community-based behavioral research. *Journal of Abnormal Child Psychology*, 2(3), 239-244.
- Wolery, M. (2013). A commentary: Single-case design technical document of the what works clearinghouse. *Remedial and Special Education*, 34(1), 39-43.
- Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, 49(2), 156-167.

Appendix A

Checklist/ Data collection tool I (1991-2019)

_____ Make sure article is from *Journal of Applied Behavior Analysis* between the years of 1991-2019. If so, continue on to opening the article.

_____ Open article and first look at the age of participants in the study. If the age/s falls between the ages of 0-19 years, the study is to be included in the research collection.

_____ Record the age of the participant in the data collection spreadsheet along with the year of publication.

_____ Search the article (hold down control + f) for key words (i.e., procedural/treatment integrity/fidelity/implementation/accuracy/independent variable implementation, etc.) If none of these words come up, read the article for mentions of integrity that may be worded other ways, such as “The protocol was followed appropriately or “the methods outlined were conducted successfully.” Determine the independent variable and look for measurements or monitoring of this variable (i.e., a checklist, monitoring video footage or observing implementation through a one-way mirror, self-monitoring, etc.).

_____ If the article provided quantitative evidence for treatment integrity such as the percentage with which the integrity was implemented---code the article as “Yes.”

_____ If the article said that they collected treatment integrity or reported a non-quantifiable measure of integrity (i.e., the protocol was implemented with high integrity) but did not provide any quantitative support for it---code the article as “Monitored.”

_____ If the article did not mention treatment integrity or stated that integrity was not collected---code the article as “no.”

Checklist/ Data collection tool II (2011-2019)

_____ Pull up the articles coded as “yes” from the years 2011-2019 and look for the setting that the study was conducted in. This is typically found in the methods section under participants and setting.

_____ Code as either school, home, clinical, or university setting. If it was conducted elsewhere then code it as other (i.e., in the community.) If it was conducted in a combination of these settings, code each location separately). List each setting/s separately for each article.

_____ Next determine who was responsible for implementing the procedures. This can typically be found in the Procedure section of the article but may be described elsewhere.

_____ Code as either Parents/Caregivers, Therapists/Clinical Staff, Teachers/Instructors, Experimenters/Researchers, Authors/Primary Investigators. If the person implementing the procedures does not fall into these categories, code at Other. If multiple implementers are listed record each one on the spread sheet separately.

Appendix B

Reliability Table: Treatment Integrity(Match or No Match: 1991-2019): *Phase I*

Year	Total # of Articles	Reliability Data Match	Total Agreement	% Agreement	Percentage of articles (20% or greater)
1991	4	4	4/4	100.0%	23.5%
1992	7	6	6/7	85.7%	22.5%
1993	6	5	5/6	83.3%	23.0%
1994	7	7	7/7	100.0%	22.6%
1995	5	5	5/5	100.0%	20.0%
1996	7	6	6/7	85.7%	20.0%
1997	8	8	8/8	100.0%	21.6%
1998	7	7	7/7	100.0%	20.0%
1999	5	4	4/5	80.0%	22.7%
2000	7	7	7/7	100.0%	20.0%
2001	7	7	7/7	100.0%	21.9%
2002	6	6	6/6	100.0%	22.2%
2003	6	6	6/6	100.0%	22.2%
2004	8	8	8/8	100.0%	22.2%
2005	7	5	5/7	71.4%	21.9%
2006	6	6	6/6	100.0%	20.0%
2007	10	9	9/10	90.0%	21.7%
2008	7	7	7/7	100.0%	20.0%
2009	11	11	11/11	100.0%	20.8%
2010	10	9	9/10	90.0%	22.2%
2011	13	12	12/13	92.3%	20.6%
2012	13	12	12/13	92.3%	21.3%
2013	10	10	10/10	100.0%	21.3%
2014	9	9	9/9	100.0%	20.5%
2015	9	9	9/9	100.0%	21.4%
2016	10	10	10/10	100.0%	20.8%
2017	8	8	8/8	100.0%	20.5%
2018	7	7	7/7	100.0%	20.0%
2019	9	9	9/9	100.0%	20.0%
Total	229	219	219/229= 95.6%		On average 21.3%/year

Appendix C

Reliability Tables: Location and Implementation Data

Reliability Table: Location:

Year	Total Number of Locations Listed	Total Number of Reliability Matches	Total Agreement	% Agreement
2011	10	9	9/10	90%
2012	7	7	7/7	100%
2013	5	4	4/5	80%
2014	7	7	7/7	100%
2015	7	7	7/7	100%
2016	7	7	7/7	100%
2017	8	7	7/8	87.5%
2018	7	6	6/7	85.7%
2019	8	7	7/8	87.5%
Totals	66	61	61/66	92.4%

Reliability Table: Implementation

Year	Total Number of Locations Listed	Total Number of Reliability Matches	Total Agreement	% Agreement
2011	9	8	8/9	88.9%
2012	6	5	5/6	83.3%
2013	5	5	5/5	100%
2014	6	6	6/6	100%
2015	6	6	6/6	100%
2016	6	6	6/6	100%
2017	6	5	5/6	83.3%
2018	7	5	5/7	71.4%
2019	7	7	7/7	100%
Totals	58	53	53/58	91.4%

Tables and Figures

Table 1. Treatment Integrity Data: (1991-2019)

Year	# of Articles	Yes	% Yes	No	% No	Monitored	% Monitored
1991	17	10	58.8%	3	17.6%	4	23.5%
1992	31	15	48.4%	12	38.7%	4	12.9%
1993	26	6	23.1%	16	61.5%	4	15.4%
1994	31	14	45.2%	13	41.9%	4	12.9%
1995	25	5	20.0%	17	68.0%	3	12.0%
1996	34	10	29.4%	22	64.7%	2	5.9%
1997	37	12	32.4%	18	48.6%	7	18.9%
1998	35	10	28.6%	20	57.1%	5	14.3%
1999	22	8	36.4%	13	59.1%	1	4.5%
2000	35	5	14.3%	26	74.3%	4	11.4%
2001	32	7	21.9%	20	62.5%	5	15.6%
2002	27	9	33.3%	17	63.0%	1	3.7%
2003	27	6	22.2%	17	63%	4	14.8%
2004	36	8	22.2%	25	69.4%	3	8.3%
2005	32	5	15.6%	23	71.9%	4	12.5%
2006	30	10	33.3%	19	63.3	1	3.33%
2007	45	10	22.2%	31	68.9%	4	8.9%
2008	33	12	36.4%	17	51.5%	4	12.1%
2009	53	15	28.3%	35	66%	3	5.6%
2010	45	14	31.1%	28	62.2%	3	6.7%
2011	63	27	42.8%	31	49.2%	5	7.9%
2012	61	24	39.3%	37	60.7%	0	0.0%
2013	47	16	34.0%	29	61.7%	2	4.2%
2014	44	19	43.2%	25	56.8%	0	0.0%
2015	42	22	52.4%	20	47.6%	0	0.0%
2016	48	26	54.2%	22	45.8%	0	0.0%
2017	39	21	53.8%	18	46.2%	0	0.0%
2018	35	17	48.6%	18	51.4%	0	0.0%
2019	45	28	62.2%	14	31.1%	3	6.7%

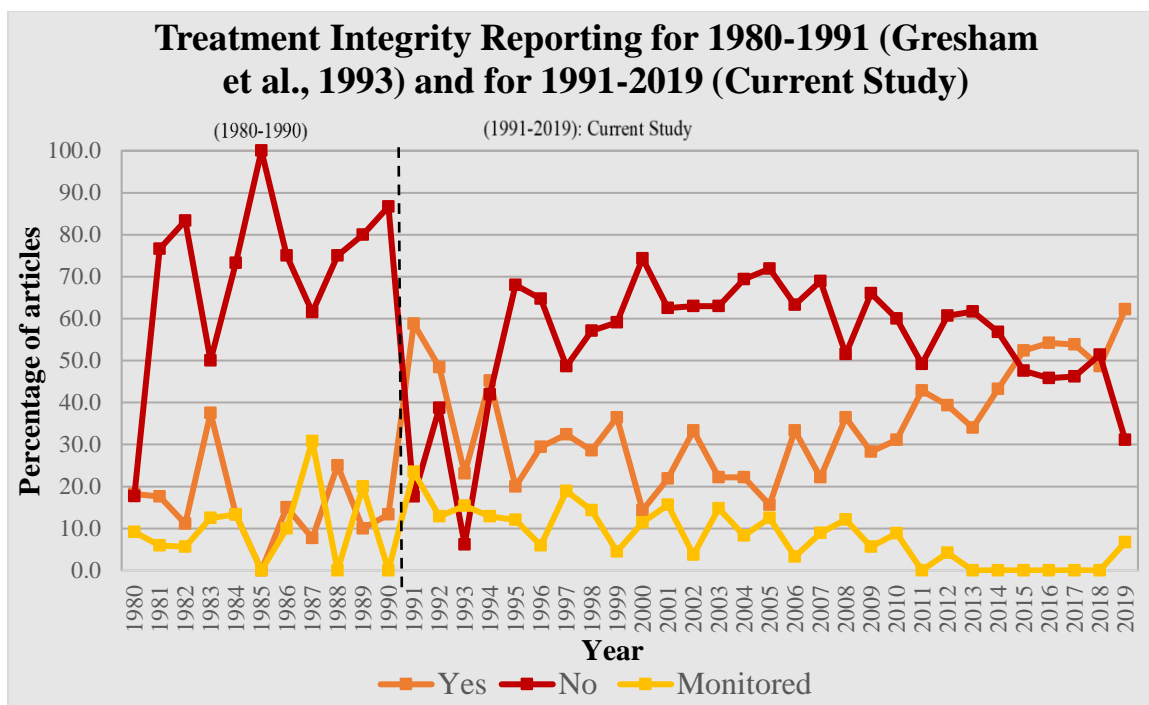


Figure 1. Continuation graph presented with a phase line between the Gresham data (1993) and the current TI data. The x-axis shows the year of publication (1980-2019) and the y-axis shows the percentage of articles. The “yes” data series represents the percentage of the total number of studies that reported integrity data. The “no” data series represents the percentage of the total number of studies that did not provide TI data. The “monitored” data series represents the percentage of the total number of studies that fell into the monitored category.

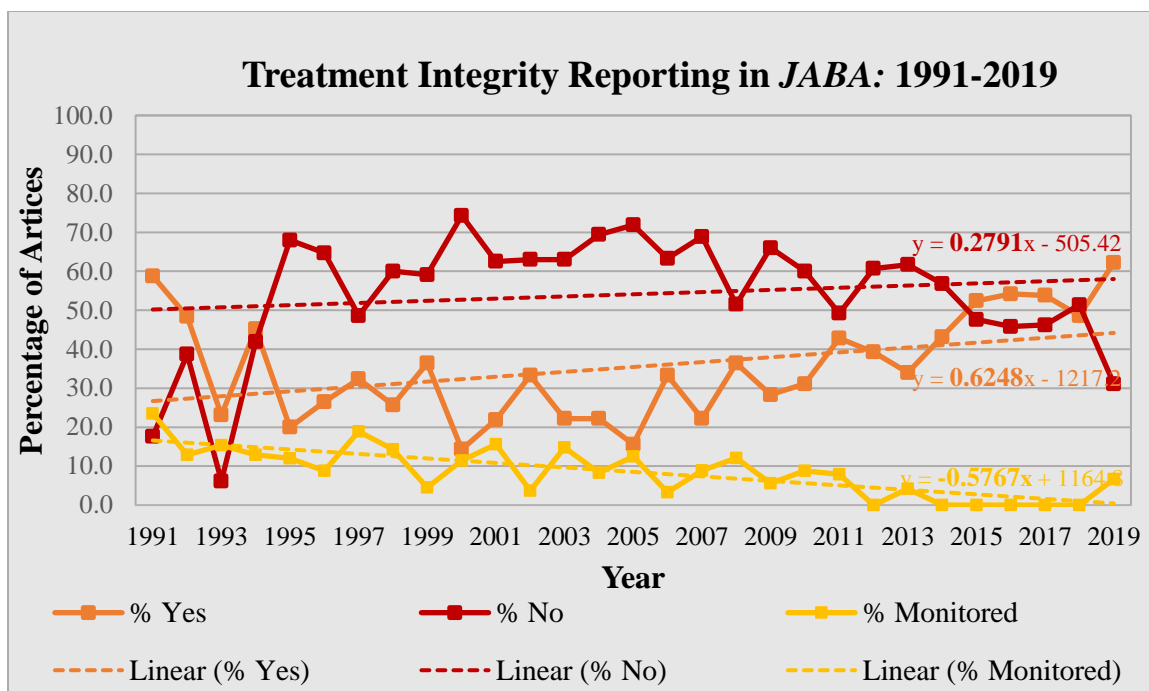


Figure 2. A visual representation of TI data from 1991-2010. The x-axis shows time (1991-2019), and the y-axis shows percentages. The “yes” data series represents the percentage of the total number of studies that reported integrity data. The “no” data series represents the percentage of the total number of studies that did not provide TI data. The “monitored” data series represents the percentage of the total number of studies that fell into the monitored category. The dotted lines corresponding to each colored data series represent the trends in integrity reporting data over this period.

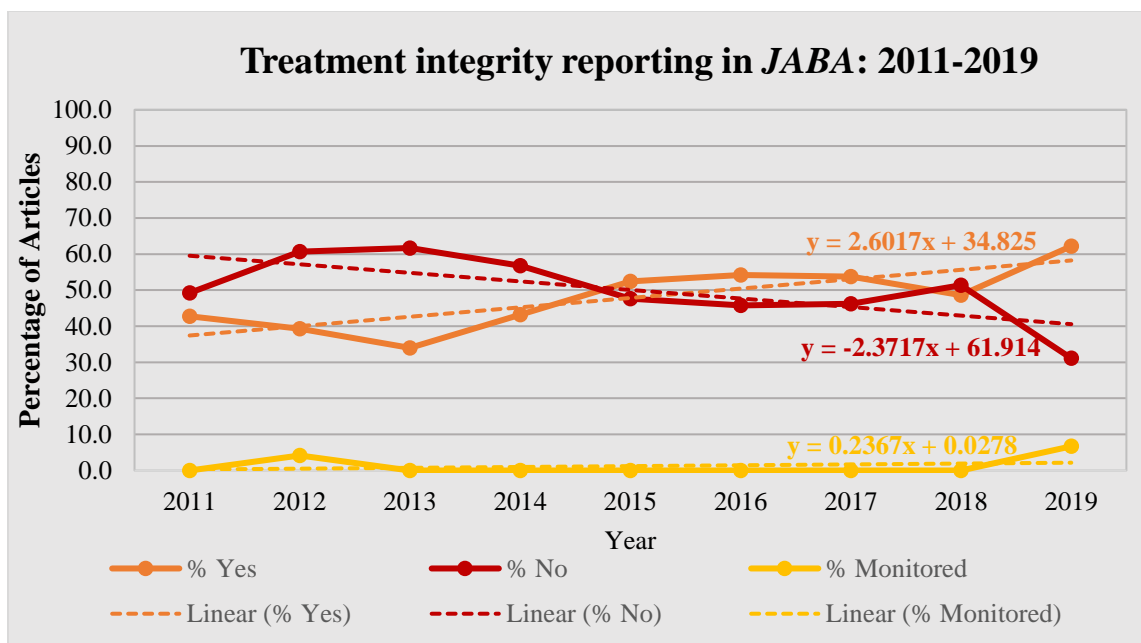


Figure 3. A closer look at TI data from the past 9 years from 2011-2019. The x-axis shows time (2011-2019), and the y-axis shows percentages. The “yes” data series represents the percentage of the total number of studies that reported integrity data. The “no” data series represents the percentage of the total number of studies that did not provide TI data. The “monitored” data series represents the percentage of the total number of studies that fell into the monitored category. The dotted lines corresponding to each colored data series represent the trends in integrity reporting data over this period.

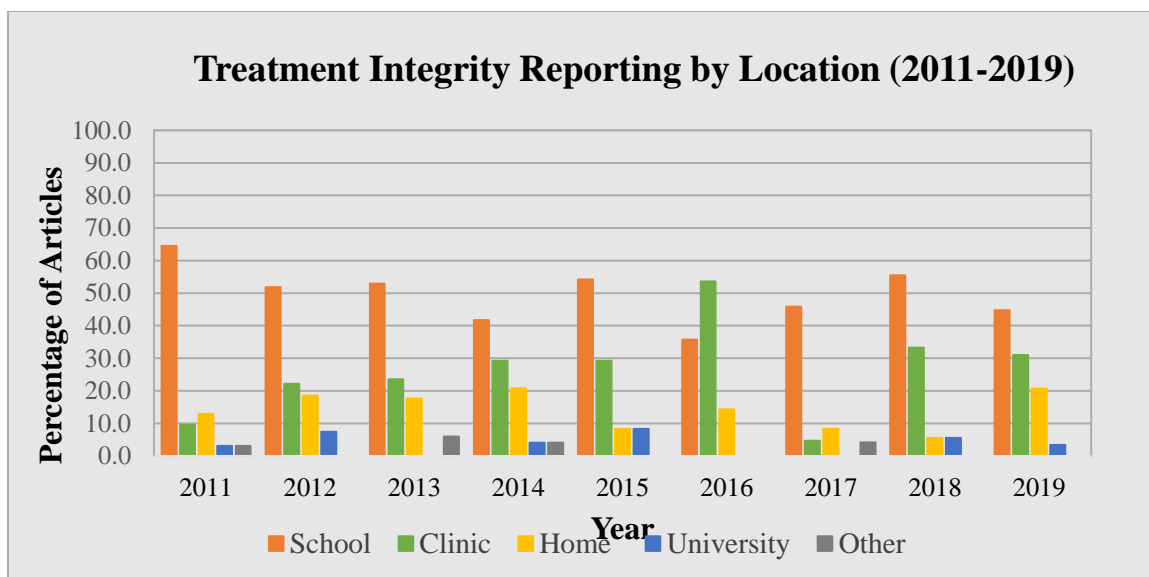


Figure 4. Chart showing the percentage of articles including TI data across the specific settings/locations where the studies were conducted. The x-axis shows the year, and the y-axis shows percentages. The orange data series represents school settings. The green data series represents clinical settings, the yellow data series represent home settings, the blue data series represents university settings, and the grey data series includes other settings (e.g., community).

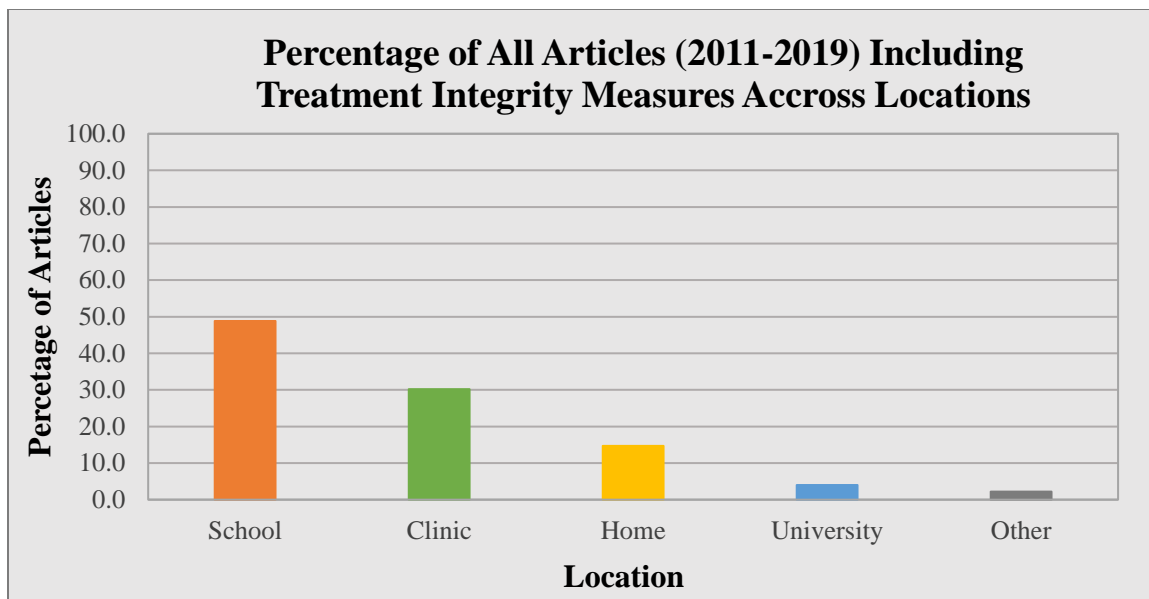


Figure 5. Chart showing the percentage of integrity data reporting across the settings/locations where the studies were conducted as a combined representation including all articles published in JABA from 2011-2019. The x-axis shows each location. The orange data series represents school settings. The green data series represents clinical settings, the yellow data series represent home settings, the blue data series represents university settings, and the grey data series includes other settings (e.g., community).

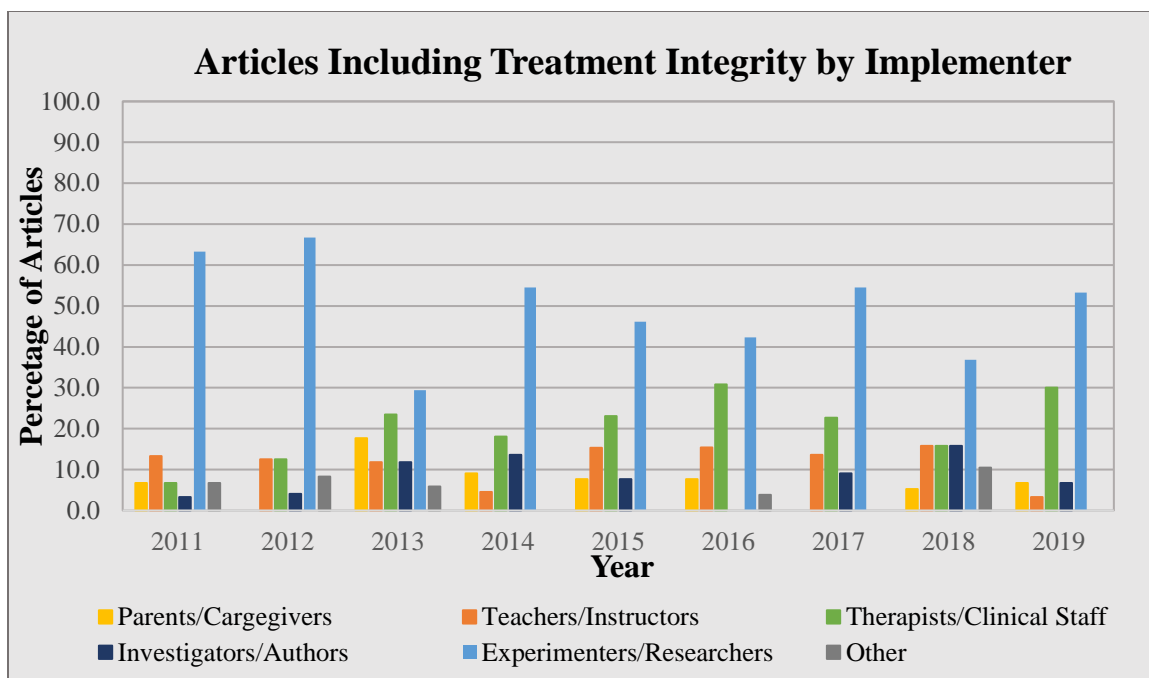


Figure 6. Graph of the percentage of integrity data reporting who implemented the procedures. They y-axis represent the percentage of articles. The x-axis shows each year. The yellow data series represents school parents/caregivers. The green data series represents therapists/clinical staff. The light blue data series represents experimenters/researchers, the orange data series represents teachers/instructors, and the grey data series includes other settings (i.e., community locations).

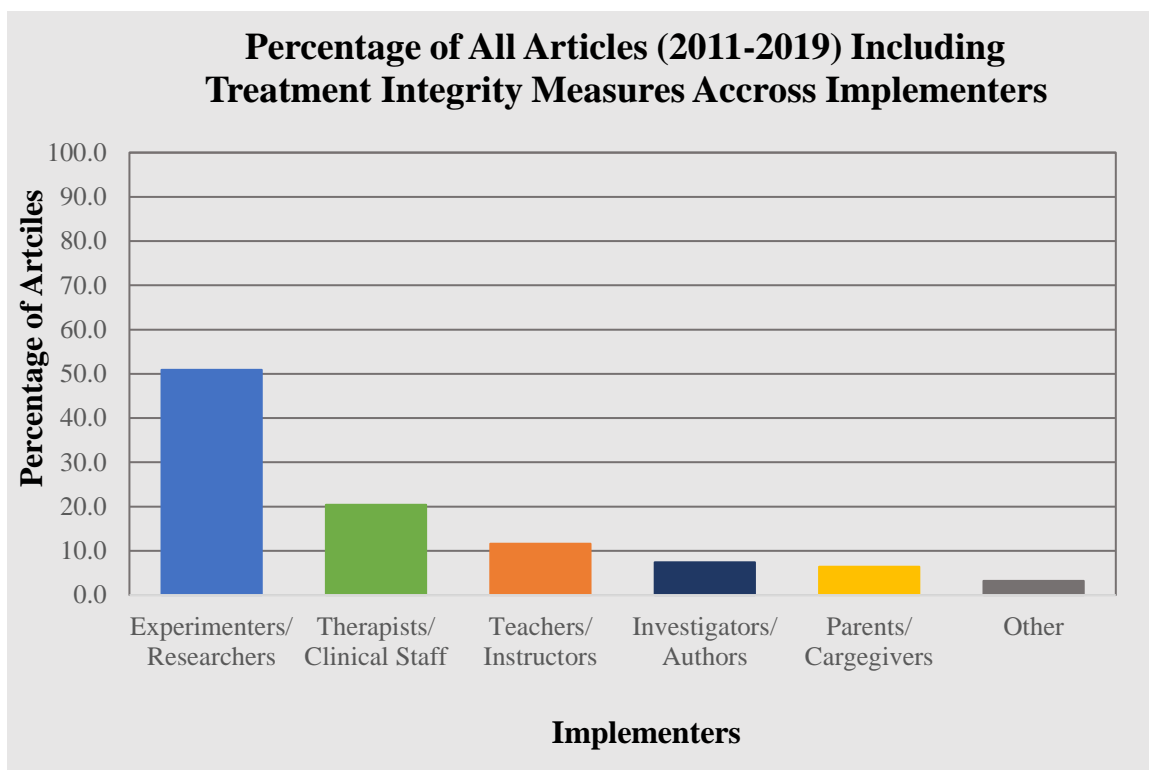


Figure 7. Percentage of integrity data reporting according to who implemented the procedures. The yellow section represents parents/caregivers. The green section represents therapists/clinical staff. The light blue data section represents experimenters/researchers, the orange section represents teachers/instructors, and the grey data series includes other settings (i.e., community).