EDUCATION AND TREATMENT OF CHILDREN Vol. 41, No. 4, 2018

# Rigor, Replication, and Reproducibility: Increasing the Relevance of Behavioral Disorders Research

Nicholas A. Gage

*University of Florida*

*Consortium for the Advancement of Special Education Research (CASPER)*

Robert N. Stevens

*South Carolina Association of Positive Behavior Supports Network*

**Abstract**

Science is being criticized within scientific domains and society. To address this, we suggest that researchers in special education, broadly, and those working in the behavioral disorders field, specifically, consider three critical facets of research: (a) importance of rigor, (b) value of replication, and (c) transparency of reproducibility in behavioral disorders. In this manuscript, we describe critical features of the three facets. Then, we provide an empirical illustration of all three research facets via a quasi-experimental design analysis of school-wide positive behavior interventions and support (SWPBIS) and its impact on discipline outcomes. Recommendations are then provided for improving the quality of behavioral disorders research.

*Keywords*: replication, reproducibility, research, behavioral disorders

Science is defined as "a system of knowledge covering truths or the operation of general laws, especially as obtained and tested through scientific method" (Merriam-Webster, 2018). More precisely, science is the cumulative knowledge discovered and tested via the scientific method. Yet recently, science and the scientific method have been criticized, both within scientific domains and broader society (Pridemore, Makel, & Plucker, 2018). Therefore, we propose that special education researchers, specifically those in the behavioral disorders field, consider approaches to address, head-on, the assault on science through rigor, replication, and reproducibility. The three

Correspondence should be addressed to: Nicholas A. Gage, School of Special Education, School Psychology, & Early Childhood Studies, University of Florida, 1403 Norman Hall, PO Box 117050, Gainesville, FL 32611, Email: gagenicholas@coe.ufl.edu

goals of this paper are to highlight (a) the importance of rigor (i.e., study quality) in behavioral disorders research, (b) the value of replication, and (c) transparency of reproducibility in order to increase the relevance of behavioral disorders research in broader educational research communities (Hedges, 2018). To meet this goal, we provide a brief overview of each facet of high quality research and an empirical example (demonstration/illustration) that encompasses each via a quasi-experimental study of the school-level effects of school-wide positive behavior interventions and supports (SWPBIS) on discipline outcomes in South Carolina.

**Rigor**

The importance of rigor in educational research is neither a novel nor new concept (cf. Campbell & Stanley, 1966; McCall, 1923). Yet, pushed by the modern evidence-based practice movement initiated by the passage of No Child Left Behind (2002) and the development of the What Works Clearinghouse (WWC), the rigor of scientific evidence in education has become widely debated (Hedges, 2018). In response, the Council for Exceptional Children (CEC) convened a task force, in 2003, to develop quality indicators for different research designs, including single-case designs and qualitative research. This broad-based approach was based on Shavelson and Towne's (2002) typography of educational research questions: (a) description (what is happening?); (b) cause (is there a systematic effect?); and (c) process or mechanism (why or how is it happening?; Odom et al., 2005). Each methodology has its own history, debates about critical features, and indicators of quality. A detailed discussion of these matters is beyond the scope of this paper. Instead, we highlight some critical features of group experimental design research that reflect high rigor. Specially, we focus on the concept of validity of research designs.

Campbell and Stanley (1966) outlined 12 factors jeopardizing the validity of group experimental research findings, and categorized those factors by internal and external validity. Eight of those factors are related to internal validity, or the extent to which an experiment rules out alternative explanations of study results (Kazdin, 2011). Internal validity refers to inferences about covariation between independent and dependent variables and whether or not the covariation is causal (i.e., causal treatment effect; Shadish, Cook, & Campbell, 2002). The eight potential confounds on internal validity are: (1) history, (2) maturation, (3) testing, (4) instrumentation, (5) statistical regression, (6) selection, (7) experimental mortality (attrition), and (8) selection-maturation interaction (see Shadish et al., 2002 for review). The remaining four factors are related to external validity. External

validity is the extent to which experimental results are generalizable or extend to variations in persons, settings, treatments, and outcomes (Kazdin, 2011; Shadish et al., 2002). Threats to external validity include: (1) reactive or interaction effect of testing; (2) interaction effects of selection biases and the experimental variable; (3) reactive effects of experimental arrangements; and (4) multiple-treatment inferences (Campbell & Stanley, 1966). Both types of validity are important for making or suggesting causality, yet internal validity is the essential condition, while external validity is never completely answerable (Shadish et al., 2002).

Rigor in group experimental research is then contingent upon how well the design of the study controls for the twelve validity confounds. The "gold standard" research design for controlling against validity confounds is the randomized controlled trial (RCT). Simply put, randomly selecting and assigning subjects to a treatment and control group results in a sample that has the same chance of receiving or not receiving the intervention. Furthermore, assuming minimal attrition, any differences between the treatment groups beyond the intervention are due to random error. However, there is debate about the ability of RCT to control against all confounds, particularly when the unit of analysis is different than the unit of assignment (i.e., analysis of student outcomes in a cluster randomized design, where schools were assigned to treatment groups or cluster randomized assignment). Recently, WWC updated their design standards for cluster randomized designs, noting that under certain circumstances, the design does not control against all potential confounds, thus reducing internal validity (WWC, 2017).

In many situations, RCT is not always possible or ethical, and researchers rely on non-randomized group experimental designs to evaluate causal questions. Under such circumstances, quasi-experimental designs (QED) are able to control against potential confounds, but they must be intentional and transparent about important sample characteristics and features in order to demonstrate rigor. Unlike in RCT, treatment and control groups are created by the researcher, either by convenience or by design limitations. To meet high quality standards (e.g., Cook et al., 2014; Gersten, et al., 2005; WWC, 2017), QED must establish baseline equivalence on the outcome measure (unless conceptually irrelevant, such as school dropout) and relevant sample characteristics (i.e. ethnicity, socioeconomic status [SES]). Equivalence is not clearly defined across all quality recommendations, but WWC (2017) defined equivalence as differences between the treatment and comparison groups of less than 0.25 standard deviation units. Groups that are equivalent or similar on the outcome measure, prior to

receiving the intervention and relevant characteristics (e.g., same percentage of low SES students in each group), permit conclusions that any differences after an intervention are the result of the intervention. However, because there are other unmeasured characteristics, results are less internally valid as those from RCT (i.e., where any differences are considered the product of random error).

As noted, a number of quality indicators and design standards have been developed to guide researchers in the development of rigorous group experimental research designs. Group experimental designs are applicable in the behavioral disorders field and are being used, but at much lower rates than other, less-rigorous group design approaches (i.e., group pre-post only designs). For example, in a review of 35 years of the journal *Behavioral Disorders*, Gage, Lewis, and Adamson (2010) found only 6% of the more than 900 published articles employed a QED or RCT design. More recently, Clarke, Zakszeski, and Kern (2018) reviewed publications in the *Journal of Positive Behavior Interventions* and found only 14% of articles used a QED or RCT design. Certainly, the behavioral disorders field has a rich history of publishing single-case designs (Gage et al., 2010), many of which are rigorous and unique to the researchers' question, but there is opportunity to increase the quantity of high quality QED and RCT studies.

The subsequent empirical example demonstrates rigor by using a quasi-experimental design approach with a treatment and comparison group of schools implementing SWPBIS. The schools have established equivalence between the treatment and comparison groups of less than 0.25 standard deviation units on the outcome measures, prior to receiving the intervention and relevant.

*Replication*

Starting in 2011, the Open Science Collaborative began a comprehensive direct replication project to reproduce 100 studies published in high impact factor psychology journals (e.g., *Psychological Science*: impact factor = 4.9). Overall, only 36% of the direct replications resulted in a statistically significant finding, and the average effect size in the replication studies was half the size of the original effect size ($d$ = 0.20 compared to $d$ = 0.40; Open Science Collaboration, 2015). Relatedly, Baker (2016), in collaboration with the journal *Nature*, surveyed 1,576 researchers in chemistry, biology, medicine and other related fields about replication in science. Of those surveyed, 90% noted that there is a slight crisis (38%) or a significant crisis (52%) of confidence due to low replicability of scientific research. The researchers noted a number of potential causes for the lack of replication or ability to

replicate original results. The three most common causes noted were selective reporting of results, pressure to publish, and low statistical power (not discussed in this paper, but a major concern in behavioral disorders research due to the limited sample of students with significant behavioral disorders [see Forness, Freeman, Paparella, Kauffman, & Walker, 2012]).

Replicating special education research is critical to ensure that a significant or meaningful research finding is repeatable, and also because identification of evidence-based practice is contingent on multiple studies conducted by different research teams using large study samples (cf. Cook et al., 2014). Yet, researchers have reported that replications are not typical practice (Travers, Cook, Therrien, & Coyne, 2016). For example, Makel and Plucker (2014) reviewed the top 100 education journals to identify how many published studies were true replications. Using the Boolean term replicat*, the authors found that researchers explicitly characterized an experiment as a replication study in only 0.1% of all published articles in those journals. Makel et al. (2016) replicated the review by Makel and Plucker using special education journals and found only 0.5% of published articles were explicit replications. Lemons et al. (2016) concurrently conducted a similar review as Makel and colleagues (2016), but focused on two different types of replication: (a) direct replication (i.e., an exact duplication), and (b) conceptual replication (i.e., duplication with an extension or modification). Lemons and colleagues (2016) found only 0.4% of articles were replication studies. However, Lemons et al. were able to compare their results with those obtained by Markel et al. (2016) and found only 15% agreement between the two reviews conducted simultaneously, suggesting difficulty in replicating a review of replication research.

At first glance, the results of these studies are concerning, especially given the importance of replications for synthesizing research and identifying evidence-based practices. However, the situation may not be as dire as it seems. Gage, Cook, and Reichow (2017) reviewed 109 meta-analyses published in special education journals and found the median number of studies included in those meta-analyses was 23. Therefore, although direct or conceptual replication is limited, studies focused on similar constructs are being conducted. Further, one issue limiting identification of replication studies is the difficulties with publishing null results (i.e., a replication with a null finding). This issue was recently addressed by Cook and Therrien (2017), who recommend that journal editors consider publishing methodologically sound null effects, particularly replication studies, and by Gage et al. (2017) who recommend that all special education meta-analyses

consider including grey literature, such as dissertations, to ensure null findings are included. We further recommend that researchers make methodologically sound null findings publicly available via websites or other research repositories if publication is not possible in order to decrease the likelihood of publication bias in special education research and improve the accuracy of meta-analytic findings (Gage et al., 2017). Nonetheless, there remains a need to address the replication crisis in special education through an increased focus on replicating existing research and increased access to null findings in research journals and other repositories.

The subsequent empirical example demonstrates replication by examining the exact same (a) treatment (i.e., SWPBIS), (b) outcome variables, and (c) covariates used in Gage, Lee, Grasley-Boy, and George (2018) and Gage, Grasley-Boy, George, Childs, and Kincaid (2018). Further, the study design and analysis procedures are also direct replications of those studies.

*Reproducibility*

Although often used synonymously, we define reproducibility differently from replication. Replication is conducting a new empirical study following the same procedures with a new sample that is identical to the original study (direct replication) or a sample or setting that is different in some form (conceptual replication). Reproducibility, as we are using it here, is the ability to reproduce the original research findings. Reproducibility has to do with the ability of independent researchers to examine the same dataset, use the same methods/code, and generate the same output/results. The key distinction is that reproducibility is reproducing the same results from the same data and analysis procedures, while replication is reproducing the same result with different data or samples. There are a number of reasons for which reproducibility is important. Nuijten, Hartgerink, van Assen, Epskamp, and Wicherts (2016) used a specialized software program to evaluate the accuracy of 250,000 *p*-values reported in major psychology journals between 1985 and 2013. The authors found that half of all published studies reported a *p*-value that was inconsistent with the test statistics and degrees of freedom. Perhaps more concerning, they found that roughly 12% of studies had *p*-values that were inconsistent with their findings and may have affected statistical conclusions. Almost all of the inconsistencies were for statistically significant *p*-values, suggesting possible bias towards reporting significant findings. Similar analyses have been conducted in experimental psychology using *p*-curve analysis (Simonsohn, Nelson, & Simmons, 2013), and in medicine using simulations of the likelihood of positive

predictive values (Ioannidis, 2005). These studies, taken together, are not suggesting that researchers are purposefully misleading or manufacturing results, but instead are concerned about the manipulation of data to obtain statistically significant results (i.e., "*p*-hacking") or drawing ad-hoc hypotheses to explain data after discovering an unexpected significant effect in an output file (i.e., hypothesizing after results are known, or HARKing; Kerr, 1998).

The best available method for evaluating the presence of reporting errors, analysis errors, or other inadvertent errors is through transparent analysis methods and opportunities for reproducing research findings. Some journals have gone so far as to request or require that researchers submit their analytic code along with manuscripts or to make data and code available upon request. For example, *Nature* requires that authors make all materials, data, code, and associated protocols promptly available to readers without undue qualifications. These requirements are not designed to question the ability or intentions of an author, but instead build upon the idea that science involves intense scrutiny from the scientific community to ensure accuracy of findings. This, in fact, is the central premise of the Center for Open Science (https://cos.io/), a group of scientists committed to increasing openness, integrity, and reproducibility of research and working with funding agencies and researchers to increase transparency and trustworthiness in science.

At this time, none of the special education research journals or journals that publish predominantly behavioral disorders research require registering with an agency such as the Center for Open Science or require making data and code available to readers. Yet, there is no reason to believe that such requirements will soon be adopted or that behavioral disorders researchers must take steps to ensure their findings are reproducible. In fact, the field has a long history of transparent interpretation of research results. Single-case design research is a completely transparent design, where all of the data is presented to the reader and he/she interprets the findings via the "inter-ocular test of statistical significance" via visual analysis (Baer, 1977). As replications of group experimental designs increase, so too, can the access of de-identified data and original statistical code from SPSS, SAS, STATA, M-Plus, or R be made available via web-based research repositories. An empirical example, data and code are available for reproducibility by any independent researcher.

**Putting It All Together**

As noted, our goal is to highlight the importance of rigor, replication, and reproducibility of research in the behavioral disorders field.

These three, interconnected facets of research are a necessary condition for the identification of evidence-based practices and for building a scientific knowledge-base that is transparent and trustworthy. Specifically, evidence-based practices should be based on research that is (a) high quality (rigor), (b) with results that have been found across multiple samples (replication), and (c) trustworthy (an independent researcher could find the same result with the same sample and analysis approach). We have completed a study that exemplifies all three facets in order to demonstrate how these facets can be enacted in a single study. Specifically, we conducted a quasi-experimental design state-level analysis on the impact of school-wide positive behavior interventions and supports on school-level discipline outcomes. In the subsequent illustration of an empirical study, we present methodology and results for a study, as described below, that is: (a) rigorous as it is designed to meet the WWC QED standards, (b) a direct replication of two prior studies (Gage, Grasely-Boy et al., 2018 & Gage, Lee et al., 2018), and (c) reproducible by making all of the data and code publicly available via Dropbox (https://www.dropbox.com/sh/171hke7qbfqi3bg/AAAxW4CxhDvRBJzjBLbtdr25a?d1=0).

**Study background.** School-wide positive behavior interventions and supports is a school-level prevention and intervention framework designed to teach pro-social behavior and prevent problem behavior using a multi-tiered approach (Sugai & Horner, 2009). Research suggests that SWPBIS has positive effects on office discipline referrals (ODR), school climate, bullying and peer victimization, organizational health, and academic achievement (Bradshaw, Mitchell, & Leaf, 2010; Gage, Leite, Childs, & Kincaid, 2017; Childs, Kincaid, George, & Gage, 2016; Horner, Sugai, & Anderson, 2010). Two studies have examined the effect of SWPBIS on discipline outcomes, including suspensions and corporal punishment, using extant state data and propensity score matching to identify baseline equivalent comparison groups in order to meet WWC QED standards. Gage, Lee et al. (2018) found statistically significant and large effects ($g$ = -0.71) for suspensions in one southeastern state, while Gage, Grasley-Boy et al. (2018) found statistically significant effects for only out-of-school suspensions ($g$ = -0.55) in another southeastern state. In this study, we aimed to replicate the significant findings for suspension in yet another southeastern state. Specific research questions for this study were as follows:

- Is there a significant and meaningful relation between SWPBIS implementation and discipline outcomes for schools that received SWPBIS training?

- Is there a significant and meaningful relation between SWPBIS implementation and discipline outcomes for schools that received SWPBIS training and implemented with fidelity?

## Method

In this study, we examined the school-level impact of SWPBIS on disciplinary outcomes in South Carolina using extant school data.

*Sample*

First, we collected all publicly available demographic data for all schools in South Carolina from the National Center for Educational Statistics' Common Core of Data for the 2013–2014 school year. Next, we collected all available data for all South Carolina public schools from the U.S. Department of Education's Civil Rights Data Collection website for all available discipline outcomes in the 2013–2014 school year (the most recent data available). Last, we obtained SWPBIS fidelity of implementation data from the director of the South Carolina SWPBIS project. We merged all available data sets, resulting in 1,499 records. However, as there are approximately 1,240 publics schools in South Carolina, we determined that some of the data sets included aggregated district data and repeated records. Therefore, we removed all schools with missing data across the three datasets. Our final analytic sample included 1,051 schools. Some schools were removed because they did not report discipline data to the Civil Rights Data Collection (e.g., virtual schools, hospital-based schools). Data were available for 135 schools that received SWPBIS training and 916 potential comparison schools. Demographic characteristics for schools trained to implement SWPBIS, schools implementing with and without fidelity, and all possible comparison schools are presented in Table 1. No data was available for the number of years the schools had been implementing SWPBIS. Schools that received SWPBIS training were more diverse than all other schools. On average, 44% of the students in trained schools were Black and 11% were Hispanic, while only 37% and 8%, respectively, in schools that did not receive SWPBIS training. In addition to demographic characteristics, we examined disciplinary outcomes prior to receiving SWPBIS training. Schools that received SWPBIS training, on average, had 40% more out-of-school suspension (OSS) incidents, 21% more in-school suspensions (ISS), and 58% more corporal punishments during the 2011–2012 school year.

**Table 1**

**Demographic Characteristics of Schools**

| | Trained (n = 135) | | Fidelity (n = 86) | | No Fidelity (n = 49) | | All Other Schools (n = 916) | | PSM Comparison Schools (n = 135) | |
|---|---|---|---|---|---|---|---|---|---|---|
| Demographic characteristics | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| School Size | 691 | 359 | 648 | 277 | 768 | 463 | 646 | 384 | 730 | 455 |
| White | 0.40 | 0.28 | 0.42 | 0.30 | 0.36 | 0.25 | 0.51 | 0.26 | 0.41 | 0.26 |
| Black | 0.44 | 0.28 | 0.43 | 0.29 | 0.46 | 0.26 | 0.37 | 0.27 | 0.45 | 0.29 |
| Hispanic | 0.11 | 0.13 | 0.10 | 0.13 | 0.12 | 0.14 | 0.08 | 0.08 | 0.10 | 0.10 |
| Free and Reduced Lunch | 0.60 | 0.23 | 0.59 | 0.23 | 0.62 | 0.21 | 0.61 | 0.22 | 0.58 | 0.23 |
| OSS 2011–2012 | 206 | 383 | 208 | 388 | 204 | 378 | 120 | 232 | 215 | 343 |
| ISS 2011–2012 | 132 | 232 | 140 | 252 | 120 | 192 | 104 | 265 | 138 | 193 |
| CP 2011–2012 | 2.10 | 4.91 | 2.14 | 4.58 | 2.04 | 5.48 | 0.90 | 3.13 | 2.39 | 4.84 |
| School Type | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Elementary | 88 | 65.2% | 60 | 69.8% | 28 | 57.1% | 529 | 57.8% | 87 | 64.4% |
| Middle | 35 | 25.9% | 21 | 24.4% | 14 | 28.6% | 193 | 21.1% | 35 | 25.9% |
| High | 12 | 8.9% | 5 | 5.8% | 7 | 14.3% | 176 | 19.2% | 13 | 9.6% |
| Other | 0 | 0.0% | 0 | 0.0% | 0 | 0.0% | 18 | 2.0% | 0 | 0.0% |

*Note.* OSS is out-of-school suspension, ISS is in-school suspension, CP is corporal punishment, PSM is propensity score matched.

*Measures*

**Fidelity of SWPBIS implementation.** One measure was used to examine implementation fidelity of SWPBIS in the 2013–2014 school year by a trained data collector from the South Carolina SWPBIS project.

***School-wide evaluation tool (SET).*** Schools assessed fidelity of implementation with the School-wide Evaluation Tool (SET; Sugai, Lewis-Palmer, Todd, & Horner, 2001). The SET is an observation and survey measure designed to assess a school's implementation of universal (Tier 1) practices. Trained data collectors spend approximately 2 hours in schools examining permanent products (e.g., discipline handbook, office discipline referral form) and interviewing students, administrators, teachers and other school staff. The SET includes 52 items and the following seven subdomains: (1) expectations defined, (2) behavioral expectations taught, (3) on-going system for rewarding behavioral expectations, (4) system for responding to behavioral violations, (5) monitoring and decision-making, (6) management, and (7) district-level support. Items are scored from "0" to "2," with "2" indicating the item is in place or schools are implementing with fidelity. The average score for each subdomain is calculated, and then the average among the subdomains is calculated, resulting in a full-scale score between 0 and 100%. Schools are considered meeting fidelity when their full-scale score is at or above 80%. Psychometric evaluations of the SET have found internal consistency of $\alpha = 0.96$, with test–retest reliability reported at $r = 0.97$, concurrent validity with the Effective Behavior Support Self-Assessment Survey at $r = 0.75$, and inter-scorer agreement of 99% (Horner et al., 2004).

**School-level covariates.** Two measures were used to examine school-level covariates.

***Student demographics.*** We included five school-level student demographic covariates in the final dataset. We captured the: (a) total student enrollment for each school, (b) the percentage of students in each school who were categorized as White, (c) the percentage categorized as Black, (d) the percentage categorized as Hispanic, and (e) the percentage of students receiving free- or reduced-lunch.

***School characteristics.*** We included three school characteristics as covariates. We captured the type of school (i.e. elementary, middle, high) and each school's longitudinal and latitudinal coordinates.

**Outcome variables.** We included three outcome variables in the final dataset: (a) in-ISS, (b) OSS, and (c) corporal punishment. The South Carolina Department of Education operationally defines each of the discipline outcomes and provides an overview of state law

pertaining to each discipline outcome (see U.S. Department of Education, 2017). We included the number of discipline outcomes reported for each school during the 2011–2012 school year and the 2013–2014 school year.

*Data Analysis*

We conducted a quasi-experimental design analysis comparing schools that received SWPBIS training to propensity score-matched (PSM) comparison schools never trained to implement SWPBIS. The procedures are a direct replication of two prior studies (Gage, Grasley-Boy et al., 2018, Gage, Lee et al., 2018).

**Propensity score matching.** Propensity score matching methods are designed to reduce bias in treatment effect estimates in experimental design studies that do not have random assignment of schools to conditions (Leite, 2017). A propensity score is the conditional probability of treatment assignment based on all available covariates (Rosenbaum & Rubin, 1983) and can be used for one-to-one matching, treatment to comparison schools. The PSM approach creates a covariate equivalent comparison group for evaluating treatment effects, meeting established standards for high quality QED research proposed by the WWC evidence standards (2014).

First, we estimated propensity scores using logistic regression and all available school-level covariates in Table 1 and each school's longitudinal and latitudinal coordinates. Then we used the estimated propensity scores to match schools using the one-to-one optimal matching method (Rosenbaum, 1989). The one-to-one optimal matching algorithm was conducted using the *matchit* (Ho, Imai, King, Stuart, & Whitworth, 2017) and *optmatch* (Hansen, Fredrickson, Fredrickson, Rcpp, & Rcpp, 2016) packages in R 3.4.1 (R Core Team, 2016). To confirm covariate equivalence, we calculated standardized mean difference effect sizes ($g$), where equivalence is defined as $g < 0.25$ standard deviations (WWC, 2014).

**Estimation of treatment effects**. All three outcome variables were all scaled as counts, therefore modeling of treatment effects relied upon Poisson regression. However, all three outcomes had very large numbers of zeros, therefore all models were estimated using zero-inflated Poisson (ZIP) regression to account for the excess zero counts (Long, 1997). Although the PSM model controlled for all available confounds on the treatment effect (Leite, 2017), we controlled for covariates with standardized mean differences greater than .05 per WWC (2014). All ZIP models were estimated using the 'pscl' package version 1.4.9 (Jackman, 2015) in R 3.4.1 (R Core Team, 2016).

**Effect sizes.** To increase interpretation of the treatment effects, we converted the treatment effect odds ratios to standardized mean difference (*g*) effect sizes, controlling for covariates with equivalence values greater than 0.05 standard deviation units, for interpretation of treatment effects based on WWC standards (2014). Conversions were conducted following procedures outlined by Borenstein, Hedges, Higgins, and Rothstein (2009).

## Results

### *Establishing Equivalence*

We used PSM to identify a covariate equivalent comparison group to evaluate the effectiveness of SWPBIS on disciplinary outcomes in South Carolina in the 2013–2014 school year. One-hundred and thirty-five schools received SWPBIS training, and, using PSM with all available covariates, we identified 135 comparison schools. Table 2 provides the equivalence statistics for all covariates. None of the effect sizes were greater than 0.025 standard deviation units, indicating that equivalence was established for the treatment and comparison schools. Based on the equivalence results, all models should control for school size, the percentage of Hispanic students, the percentage of students receiving free- or reduced-lunch, and corporal punishment incidents in 2011–2012. For models that examine differences by fidelity, all covariates should be included, because many of the covariate effect sizes were greater than 0.05 standard deviations.

### *Treatment Effects*

We estimated six ZIP models to evaluate the effect of SWPBIS on disciplinary outcomes using the PSM sample. The first three models are presented in Table 3 and evaluate whether there was a significant effect for schools that received SWPBIS training. As noted above, the models included all covariates with equivalence >0.05 standard deviation units. Across all three discipline outcomes, there was not a significant effect, meaning there was a difference in the frequency of ISS, OSS, or corporal punishment in schools trained to implement SWPBIS and PSM comparison schools. Next, we modeled fidelity level, all available covariates, and disciplinary outcomes (tables available from the first author) and found similar results. Overall, no significant treatment effect was found across all six models.

**Table 2**
**Sample Equivalence Effect Sizes**

| School Characteristic | Trained x PSM | Fidelity x PSM | No Fidelity x PSM |
|---|---|---|---|
| School Size | -0.09* | -0.21* | 0.08* |
| White | -0.02 | 0.06* | -0.17* |
| Black | -0.02 | -0.07* | 0.07* |
| Hispanic | 0.11* | 0.04 | 0.25* |
| Free and Reduced Lunch | 0.10* | 0.05 | 0.18* |
| Elementary | 0.02 | 0.15* | -0.19* |
| Secondary | -0.03 | -0.13* | 0.20* |
| OSS 2011–2012 | -0.02 | -0.02 | -0.03 |
| ISS 2011–2012 | -0.03 | 0.01 | -0.09* |
| CP 2011–2012 | -0.06* | -0.05 | -0.07* |

*Note*. OSS is out-of-school suspension, ISS is in-school suspension, CP is corporal punishment, PSM is propensity score matched. Statistics are all standardized mean difference effect sizes. Equivalence is defined as effect sizes < 0.25 standard deviation units. *All models should adjust for effect sizes >0.05 and < 0.25.

**Table 3**
**Zero-Inflated Poisson Regression Models for Suspensions and Corporal Punishment**

| | ISS | | OSS | | Corporal Punishment | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | se | Estimate | se | Estimate | se |
| Intercept | -0.36 | 0.88 | 1.18 | 1.5 | 1.61 | 5.58 |
| SWPBIS | 0.17 | 0.34 | -0.95 | 0.72 | -0.42 | 2.11 |
| School Size | -0.01*** | 0.00 | -0.00* | 0.00 | 0.00 | 0.00 |
| Hispanic | 1.58 | 1.27 | 0.47 | 2.96 | 11.67 | 17.91 |
| Free or Reduced Lunch | 0.87 | 0.91 | -3.19 | 1.71 | 6.86 | 9.01 |
| CP 2011–2012 | -0.36* | 0.16 | -0.09 | 0.10 | -1.11** | 0.40 |

*Note.* CP is corporal punishment, OSS is out-of-school suspension, ISS is in-school suspension, SWPBIS is a dichotomous indicator for schools trained to implement school-wide positive behavior intervention and supports; OSS is out-of-school suspensions; CP is corporal punishment. $p < 0.000$ ***, $p < 0.01$ **, $p < * 0.05$.

*Effect Sizes*

Although there was not a significant treatment effect during the 2013–2014 school year, we calculated effect sizes and their variances for later meta-analysis opportunities. First, we converted all three disciplinary outcomes to per student rates by dividing each outcome by school size. Next, we estimated marginal means for all outcomes by treatment status using a general linear model and including all covariates to ensure the effect sizes met WWC standards. The average frequency of each discipline outcome, marginal mean for the rate of each outcome variable, standard deviation of the rates, standardized mean difference effect sizes ($g$), and the variance for each $g$ are presented in Table 4. All effect sizes were small, and none approached clinical significance (0.25 standard deviation units).

## Discussion

The purpose of this conceptual paper was to provide an empirical example of rigor, replication, and reproducibility. We used a QED approach, established baseline equivalence by identifying a matched comparison group using PSM, conducted a series of analyses and made our code publicly available, and calculated effect sizes for future meta-analyses. That being said, the results are worth further discussion.

Overall, we did not find a significant or meaningful effect for schools trained to implement SWPBIS or for schools implementing with fidelity. There are a number of possible reasons for the limited effect. During the 2013–2014 school year, South Carolina was still in the early stages of scaling up effective SWPBIS training and support. At that time, training was not centralized and consistent, but regionally supported by district and regional staff who were also still beginning to develop their SWPBIS expertise. Further, there was limited statewide leadership and limited consistent follow-up by qualified SWPBIS coaches. Lastly, as noted above, no data were available for the number of years each school had been implementing. Anecdotally, most of the included schools were in their first year of implementation. Gage, Grasley-Boy et al. (2018) found that effects were much stronger for schools with 3–5 years of experience implementing SWPBIS. After more up-to-date suspension and corporal punishment data become available, a replication study could examine whether experience and improved state-wide training indeed have an effect on discipline outcomes.

**Table 4**
**Treatment Effects by Training and Fidelity Levels**

| | | SWPBIS Schools | | | PSM Comparison ($n$ = 135) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Frequency | *M* | *SD* | Frequency | *M* | *SD* | *g* | Variance of *g* |
| Trained ($n$ = 135) | | | | | | | | | |
| | OSS | 219 | 0.71 | 2.03 | 220 | 0.51 | 1.45 | 0.12 | 0.028 |
| | ISS | 136 | 0.24 | 0.45 | 149 | 0.22 | 0.31 | 0.07 | 0.028 |
| | CP | 1.24 | 0.00 | 0.01 | 1.64 | 0.00 | 0.01 | -0.09 | 0.028 |
| Fidelity ($n$ = 86) | | | | | | | | | |
| | OSS | 222 | 0.77 | 2.25 | 220 | 0.51 | 1.45 | 0.15 | 0.019 |
| | ISS | 133 | 0.26 | 0.53 | 149 | 0.22 | 0.31 | 0.11 | 0.019 |
| | CP | 1.4 | 0.00 | 0.01 | 1.64 | 0.00 | 0.01 | 0.00 | 0.019 |
| No Fidelity ($n$ = 39) | | | | | | | | | |
| | OSS | 213 | 0.75 | 2.17 | 220 | 0.51 | 1.45 | 0.13 | 0.015 |
| | ISS | 142 | 0.25 | 0.50 | 149 | 0.22 | 0.31 | 0.09 | 0.015 |
| | CP | 0.98 | 0.00 | 0.01 | 1.64 | 0.00 | 0.01 | -0.03 | 0.015 |

*Note.* OSS is out-of-school suspension, ISS is in-school suspension, CP is corporal punishment, PSM is propensity score matched, SWPBIS is a dichotomous indicator for schools trained to implement school-wide positive behavior intervention and supports.

Although the empirical example was designed to meet the three facets of research, a number of limitations necessitate mentions. First, we did not have the number of years implementing SWPBIS available in the dataset. As noted, years of experience is an important predictor of SWPBIS effectiveness. Second, we only had fidelity of implementation data available for the universal tier of implementation, and thus we cannot confirm if schools in the treatment group also implemented Tier 2 or Tier 3 practices with fidelity. Third, we had no information about any behavioral initiatives in any of the comparison schools that may have addressed or impacted discipline outcomes. Fourth, there are many additional confounding factors that may have impacted differences by treatment condition that were not included in the PSM. For example, we did not include academic achievement or a measure of principal leadership. Last, the ZIP modeling framework does not adjust for district-level cluster. To ensure accuracy of findings, we conducted a mixed effects model accounting for district-level clustering with rates of discipline outcomes and corroborated the results (model code included with all other code). We did not report them because the rates were also negatively skewed and did not meet the assumption of normality. These limitations are similar to those in Gage, Grasley-Boy et al. (2018) and Gage, Lee et al. (2018).

*Recommendations for Rigor, Replication, and Reproducibility*

As noted, rigor, replication, and reproducibility are necessary conditions for the identification of evidence-based practices. Although the purpose of this paper was not to review the presence of each of these facets in the behavioral disorders field, we believe there is opportunity to increase each facet. As such, we make the following recommendations. First, researchers and doctoral students need training and practice in developing, conducting, and reviewing rigorous research designs using established quality indicators (e.g., WWC, CEC). University courses and conference-based workshops should be developed to highlight the indicators of high quality, rigorous research across myriad research designs (e.g., single-case, qualitative). Journal editors and reviewers should also be well-versed in quality indicators and should hold researchers to a high standard and, when those standards are not met, have the authors be explicit and transparent about why indicators were not met.

Second, the field of special education and, specifically, behavioral disorders researchers, should commit to conducting direct and conceptual replications of practices that have established treatment effects and significant relationships. Replication in single-case research is a central feature of the method, both within and across studies

(Kazdin, 2011), but the same is not true in correlational and group experimental research. Certainly, with regards to group experimental research, replication can be cost prohibitive. Nonetheless, direct and conceptual replications are needed and can provide excellent opportunities for doctoral dissertations or projects for research methods courses. Third, researchers should consider making their statistical code and de-identified data available via the Center for Open Science, websites, or by invitation in manuscripts to readers. We are not recommending that journal editors not publish studies until after they have reproduced authors' results, which has been suggested (Colaresi, 2016). Instead, we suggest authors be proactive in case they, or others, have an interest in reproducing their results. Although saving code for reproducibility may not be typical practice for many researchers, we believe that learning how to do so increases the transparency and trustworthiness of research findings.

*Conclusion*

Science is experiencing heightened scrutiny and attack from both the general public and within scientific domains. Scrutiny is a valuable and necessary component of doing "good science". This is the fundamental philosophy of science developed by Popper (1959), that scientific discovery is about falsification, not proof. We contend that through rigor, replication, and reproducibility, special education and behavioral disorders research can be exemplars of transparent and trustworthy research. The model is part of our rich single-case design history and can and should be embraced across research methodologies. In this paper, we have attempted to provide an example for how to move forward, to a community committed to the three pillars of high quality research and necessary elements in the search for evidence-based practice.

## References

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis, 10*, 167–172.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News, 533*, 452–454. doi:10.1038/533452a

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons.

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized

controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions, 12*, 133–148. doi: 10.1177/1098300709334798

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Childs, K. E., Kincaid, D., George, H. P., & Gage, N. A. (2016). The relationship between school-wide implementation of positive behavior intervention and supports and student discipline outcomes. *Journal of Positive Behavior Interventions*, *18*, 89–99. doi: 10.1177/1098300715590398

Clarke, S., Zakszeski, B. N., & Kern, L. (2018). Trends in JPBI publications, 1999–2016. *Journal of Positive Behavior Interventions*, *20*, 6–14. doi: 10.1177/1098300717722359

Colaresi, M. (2016). Preplication, replication: A proposal to efficiently uprade journal replication standards. *International Studies Perspective*, *17*, 367–378.

Cook, B., Buysse, V., Klingner, J., Landrum, T., McWilliam, R., Tankersley, M., & Test, D. (2014). Council for Exceptional Children: Standards for evidence-based practices in special education. *Teaching Exceptional Children*, *46*, 206–212.

Cook, B. G., & Therrien, W. J. (2017). Null effects and publication bias in special education research. *Behavioral Disorders*, *42*, 149–158. Doi: 10.1177/0198742917709473

Forness, S. R., Freeman, S. F., Paparella, T., Kauffman, J. M., & Walker, H. M. (2012). Special education implications of point and cumulative prevalence for children with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders*, *20*, 4–18. Doi: 10.1177/1063426611401624

Gage, N. A., Cook, B., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, *83*, 428–445. doi: 10.1177/0014402917691016

Gage, N. A., Grasley-Boy, N., George, H. P., Childs, K., & Kincaid, D. (2018). A quasi-experimental design analysis of the effects of school-wide positive behavior interventions and supports on discipline in Florida. *Journal of Positive Behavior Interventions.* Advance online publication. doi: 10.1177/1098300718768208

Gage, N. A., Lee, A., Grasley-Boy, N., & George, H. P. (2018). The impact of school-wide positive behavior interventions and supports on school suspensions: A state-wide quasi-experimental analysis. *Journal of Positive Behavior Interventions.* Advance online publication. doi: 10.1177/1098300718768204

Gage, N. A., Leite, W. L., Childs, K., & Kincaid, D. (2017). Average treatment effect of school-wide positive behavior supports (SWPBIS) on school-level academic achievement in Florida. *Journal of Positive Behavior Interventions*, *19*, 158–167. doi. 10 .1177/1098300717693556

Gage, N. A., Lewis, T. J., & Adamson, R. (2010). Where have we been, where are we going?: 35 years of behavioral disorders. *Behavioral Disorders*, *35*, 280–293.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Hansen, B.B., & Klopfer, S.O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Studies, 15*, 609–627. doi: 10.1198/106186006X137047

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*, 1–21. doi: 10.1080/19345747.2017.1375583

Ho, D., Imai, K., King, G., Stuart, E., & Whitworth, A. (2017). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*, 1–28.

Horner, R. H., Sugai, G., & Anderson, C. M. (2010). Examining the evidence base for school-wide positive behavior support. *Focus on Exceptionality, 42*, 1–14.

Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Boland, J. B. (2004). The school-wide evaluation tool (SET) a research instrument for assessing school-wide positive behavior support. *Journal of Positive Behavior Interventions*, *6*, 3–12.

Ioannidis J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, 696–701. doi:10.1371/journal.pmed.0020124

Jackman, S. (2015). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. R package version 1.4.9. URL http://pscl .stanford.edu/

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Leite, W. (2017). *Propensity score methods using R*. Los Angeles, CA: Sage.

Lemons, C. J., King, S. A., Davidson, K. A., Berryessa, T. L., Gajjar, S. A., & Sacks, L. H. (2016). An inadvertent concurrent replication: Same roadmap, different journey. *Remedial and Special Education*, *37*, 213–222. Doi: 10.1177/0741932516631116

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage Publications.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316.

Makel, M. C., Plucker, J. A., Freeman, J., Lombardino, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research: Necessary but far too rare. *Remedial and Special Education*, *37*, 205–2012. 10.1177/0741932516646083

McCall, W. A. (1923). *How to experiment in education*. New York, NY: Macmillan.

No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. § 6319 (2002).

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidenced-based practices. *Exceptional Children*, *71*, 137–149.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* 943–950. doi: 10.1126/science .aac4716

Popper, K. (1959, 2002). *The logic of scientific discovery.* New York: Routledge.

Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology*, *1*, 19–38. doi: 10.1146/annurev-criminol-032317–091849

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from https://www.R-project.org/

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*, 1024–1032.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Science [Def. 1]. (n.d.). *Merriam-Webster Online.* Retrieved from http://www.merriamwebster.com/dictionary/citation

Shadish, W. R., Cook, T. D., & Campbell, T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology*, 143, 534–547. http://dx.doi.org/10.1037/a0033242

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality*, *17*, 223–237.

Sugai, G., Lewis-Palmer, T. L., Todd, A. W., & Horner, R. H. (2001). School-wide evaluation tool (SET). *Eugene, OR: Center for Positive Behavioral Supports, University of Oregon*, 94–101.

Travers, J. C., Cook, B. G., Therrien, W. J., & Coyne, M. D. (2017). Replication research and special education. *Remedial and Special Education*, *37*, 195–204. doi: 10.1177/0741932516648462

What Works Clearinghouse. (2014). What Works Clearinghouse procedures and standards handbook (Version 3.0). Washington D.C. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

What Works Clearinghouse. (2017). WWC cluster design standards. Washington D.C. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_cluster_standards_030416.pdf

U.S. Department of Education. (2017). South Carolina compilation of school discipline laws and regulations. Washington D.C. Retrieved from https://safesupportivelearning.ed.gov/sites/default/files/disciplinecompendium/South%20Carolina%20School%20Discipline%20Laws%20and%20Regulaions.pdf