

# Promoting Positive Behavior Using the Good Behavior Game: A Meta-Analysis of Single-Case Research

Journal of Positive Behavior Interventions  
1-11  
© Hammill Institute on Disabilities 2015  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1098300715592355  
jpbj.sagepub.com  


Lisa Bowman-Perrott, PhD<sup>1</sup>, Mack D. Burke, PhD<sup>1</sup>,  
Samar Zaini, MS<sup>1</sup>, Nan Zhang, MA<sup>1</sup>, and Kimberly Vannest, PhD<sup>1</sup>

## Abstract

The Good Behavior Game (GBG) is a classroom management strategy that uses an interdependent group-oriented contingency to promote prosocial behavior and decrease problem behavior. This meta-analysis synthesized single-case research (SCR) on the GBG across 21 studies, representing 1,580 students in pre-kindergarten through Grade 12. The TauU effect size across 137 phase contrasts was .82 with a confidence interval 95% CI = [0.78, 0.87], indicating a substantial reduction in problem behavior and an increase in prosocial behavior for participating students. Five potential moderators were examined: emotional and behavioral disorder (EBD) risk status, reinforcement frequency, target behaviors, GBG format, and grade level. Findings suggest that the GBG is most effective in reducing disruptive and off-task behaviors, and that students with or at risk for EBD benefit most from the intervention. Implications for research and practice are discussed.

## Keywords

behavior(s), challenging, intervention(s), single-case designs, meta-analysis, studies

There is a need to identify effective prevention-oriented approaches to behavior management, especially at the classroom level (Simonsen, Fairbanks, Briesch, Myers, & Sugai, 2008). The Good Behavior Game (GBG) is a universal behavior management strategy that uses an interdependent group-oriented contingency to promote positive classroom behaviors. First introduced by Barrish, Saunders, and Wolf (1969), the GBG was originally developed to reduce disruptive behaviors in an elementary school classroom. More recently, it has been used within prevention science (Kellam, Rebok, Ialongo, & Mayer, 1994). Embry (2002) referred to the GBG as a “behavioral vaccine” based, in part, on seminal research conducted by Kellam et al. (1994). The authors reported positive long-term impacts of the intervention on aggressive and disruptive behaviors from a large-scale epidemiological trial.

The major features of the GBG as described by Barrish et al. (1969) included the following: (a) assigning students to teams, (b) giving points to teams that exhibit inappropriate behaviors, and (c) rewarding the team that accumulated the lowest number of points (i.e., the team that exhibits the least amount of problem behavior). Depending on how the GBG is set up, more than one team can win if the criterion for winning (e.g., five or fewer points) is reached. In some instances, the GBG has been modified as follows: (a) rewarding appropriate behaviors (Crouch, Gresham, & Wright, 1985), (b)

adding a merit system for simultaneously promoting academic engagement (Darveaux, 1984), (c) adding a behavioral intervention (Wright & McCurdy, 2011), (d) including a self-monitoring component (Babyak, Luze, & Kamps, 2000), (e) examining the impact of not using teams (Harris & Sherman, 1973), (f) investigating the effect of using independent and dependent (vs. interdependent) group contingencies (Gresham & Gresham, 1982), and (g) allowing individual students to earn points (Babyak et al., 2000).

The GBG is effective across a variety of problem behaviors including verbal and physical aggression (Saigh & Umar, 1983), noncompliance (Swiezy, Matson, & Box, 1992), oppositional behaviors (Leflot, van Lier, Onghena, & Colpin, 2010), hyperactive behaviors (Huizink, van Lier, & Crijnen, 2008), and out-of-seat behaviors (Medland & Stachnik, 1972). Increases in prosocial behaviors associated with the Game include on-task behaviors (Rodriguez,

<sup>1</sup>Texas A&M University, College Station, USA

## Corresponding Author:

Lisa Bowman-Perrott, Texas A&M University, 4225 TAMU, College Station, TX 77843-4225, USA.  
Email: lbperrott@tamu.edu

**Action Editor:** Kathleen Lane

2010), assignment completion (Darveaux, 1984), acceptance of authority (Dolan et al., 1993), and improved concentration (Dolan et al., 1993). Furthermore, positive outcomes from participation in the GBG have been observed in both general education (McGoey, Schneider, Rezzetano, Prodan, & Tankersley, 2010) and special education settings (Salend, Reynolds, & Coyle, 1989). Finally, although most of the research on the GBG has been conducted in elementary schools (Ruiz-Olivares, Pino, & Herruzo, 2010), there is promising evidence of its efficacy with middle and high school students (Kleinman & Saigh, 2011).

## Previous GBG Reviews

Two GBG literature reviews and one meta-analysis have been published to date (Flower, McKenna, Bunuan, Muething, & Vega, 2014; Tankersley, 1995; Tingstrom, Sterling-Turner, & Wilczynski, 2006). Tankersley (1995) reviewed nine studies published between 1969 and 1994, including six single-case research (SCR) studies. Studies focused solely on elementary school students, with one study (Kellam et al., 1994) following up with participants in middle school. Of the nine studies, seven used the original GBG format described by Barrish et al. (1969). The remaining two studies (Fishbein & Wasik, 1981; Harris & Sherman, 1973) implemented modified versions of the Game by reinforcing positive behaviors. Tankersley (1995) reported that (a) the GBG was effective in reducing problem behaviors, (b) improvements in academic engagement could be attributed to the GBG (Darveaux, 1984), (c) social validity was high among teachers and students (Salend et al., 1989), and (d) direct observations were primarily used, with the exception of two studies which included teacher ratings of student behaviors (Kellam et al., 1994) and/or peer nominations (Dolan et al., 1993).

Tingstrom et al. (2006) synthesized 29 studies published between 1969 and 2000, 21 of which used SCR designs. Twenty of the 29 studies focused on elementary school students, 2 examined outcomes for middle and high school students, and 4 included both elementary and secondary students. One study did not report students' grade level; the remaining two studies did not investigate student outcomes. Most of the participants were "students of typical development in general education classes or students with a history of behavior problems" (Tingstrom et al., 2006, p. 241). Only a few studies examined the efficacy of the GBG with students with disabilities. In addition, studies were conducted in the United States, Germany (Huber, 1979), and Sudan (Saigh & Umar, 1983). Tingstrom et al. reported findings consistent with those of Tankersley's (1995) review. Moreover, Tingstrom et al. noted that the GBG was effective regardless of whether (a) the criterion for reinforcement was changed (Harris & Sherman, 1973) or (b) the original format described by Barrish et al. (1969) or variations of it were used (e.g., Swiezy et al., 1992). Although the literature reviews provide valuable information about the GBG, there

are several limitations. First, neither reported effect sizes with confidence intervals (CIs); CIs are needed for accurately interpreting effect size data (Cooper, 2011; Thompson, 2007). Second, although they summarized several variables (e.g., reinforcement frequency, GBG format), it is not known which is more effective in promoting positive behaviors. Third, although both reviews identified the need to examine outcomes for students with disabilities, neither summarized data for this group of students. Yet, Salend et al. (1989), for example, implemented the GBG with students with emotional and behavioral disorders (EBD).

Flower et al. (2014) examined the impact of the GBG on challenging behaviors across 22 studies, 16 of which used SCR designs. They examined the impact of fidelity of implementation, training for interventionists, intervention duration, setting, and the use of rewards on problem behaviors. The authors found that although few studies reported fidelity, only one reported low fidelity. They indicated that a lower fidelity rating in that study did not minimize the benefit students gained from participating in the GBG. They concluded that (a) results did not seem to be affected by the type of interventionist training, (b) shorter interventions produced change in student behavior, (c) elementary and secondary students demonstrated a reduction in problem behaviors, and (d) the use of rewards had a positive impact on decreasing inappropriate behaviors while increasing appropriate behaviors, particularly when students found rewards reinforcing.

The Flower et al. (2014) meta-analysis extends the contribution of the literature reviews. However, several important considerations remain unaddressed. First, an investigation of the following is needed: (a) whether the GBG has differential effects for students with or at risk for EBD (as they characteristically demonstrate higher rates of challenging behavior; Reid, Gonzalez, Nordess, Trout, & Epstein, 2004), (b) whether the frequency of reinforcement moderates student outcomes, and (c) whether GBG format affects student outcomes. Second, Flower et al. only included studies published in peer-reviewed journals. As they noted, "sound implementations of the GBG conducted for dissertation . . . research may have been missed" (p. 20). Third, a measure of design quality (e.g., What Works Clearinghouse [WWC]; Kratochwill et al., 2010) was not included. Fourth, the reporting of effect sizes with CIs is missing. Given these limitations, a study that addresses these gaps and further extends the literature is essential to better understand the impact of this widely used intervention.

## Purpose of the Study and Research Questions

Often, studies using SCR designs are excluded from meta-analyses (Allison & Gorman, 1993). Perhaps this is because recommended standards for quality SCR designs and evidence of treatment effects have only recently been disseminated

(Kratochwill et al., 2010). Quantitative syntheses are critical for establishing the evidence-base for effective behavioral interventions and practices (Parker & Hagan-Burke, 2007), especially with regard to SCR (Shadish, Rindskopf, & Hedges, 2008). Meta-analysis “allows researchers to arrive at conclusions that are more accurate and more credible than can be presented in any one primary study or in a non-quantitative, narrative review” (Rosenthal & DiMatteo, 2001, p. 61). The purpose of the current meta-analysis was to quantitatively analyze the SCR literature on the GBG to examine its impact: (a) for students with or at risk for EBD, (b) with regard to reinforcement frequency, (c) across target behaviors, (d) based on the GBG format used, and (e) across grade levels. Two main research questions were addressed:

**Research Question 1:** What is the overall effect of the GBG across studies?

**Research Question 2:** What are the effects of potential moderators on students’ behavioral outcomes?

## Method

### *Literature Search, Inclusion Criteria, and Design Quality*

To identify relevant studies, a search of the literature was conducted using the Education Full Text, Educational Resources Information Center (ERIC), PsycINFO, and Dissertations and Theses Full Text databases. Dissertations and unpublished studies were sought for inclusion to help reduce the possibility of publication bias, the tendency for only studies yielding favorable results to be published (Rosenthal & DiMatteo, 2001). To identify the maximum number of potentially eligible studies, we used the term *Good Behavior Game*; 272 search results were obtained. The first author and two doctoral students reviewed titles and abstracts for relevance; articles were reviewed when more information was needed. Articles were excluded if they (a) included some combination of the search terms but were unrelated, (b) used a group design (as most GBG studies used SCR designs), (c) were literature reviews, (d) were duplicate studies, or (e) were studies for which a complete article copy was unavailable (e.g., older studies). Thirty-four studies remained. Studies that investigated non-behavioral outcomes or focused on teacher outcomes but did not investigate student outcomes were excluded as well, resulting in 24 studies. We also conducted an ancestral search for studies in the references of articles identified in the electronic database search; no additional studies were found. To be included, studies had to (a) implement the GBG to reduce problem behavior or increase appropriate behavior, (b) involve participants in pre-kindergarten through Grade 12, (c) be published in a peer-reviewed journal or conducted as dissertation research or an unpublished article between

1969 and 2013, (d) use an SCR design, (e) provide graphed data of student outcomes, and (f) be reported in English. To ensure that basic quality standards were adhered to from the initial pool of studies, we applied two WWC (Kratochwill et al., 2010) standards to identify studies that used a design that (a) could demonstrate experimental control (viz., reversal, multiple baseline) and (b) had at least three data points per phase. The application of these criteria yielded 21 studies for inclusion in this meta-analysis.

We then used a rubric adapted from Maggin, Chafouleas, Goddard, and Johnson (2011) to evaluate the included studies across four WWC SCR design standards (see Table 1). First, we determined whether the GBG was systematically manipulated. Second, we determined whether the design could demonstrate an experimental effect across three points in time or with three phase changes. Third, we evaluated studies using reversal designs ( $n = 10$ ) to ensure that they had a minimum of four phases with at least five data points per phase to meet standards, or at least three data points per phase to meet standards with reservations. Studies using multiple baseline designs ( $n = 11$ ) were evaluated to ensure that they included at least six phases and at least five data points per phase to meet standards, and three data points per phase to meet standards with reservations. Fourth, we examined each study for interobserver agreement (IOA). Based on these four standards, each study was categorized as *meets standards*, *meets standards with reservations*, or *does not meet standards* (Kratochwill et al., 2010). With regard to design quality, 5 of the 21 studies met standards, 9 met standards with reservations, and 7 did not meet standards (primarily because of missing IOA data on the percentage of observations; see Table 1). To establish whether there was evidence of an effect, we conducted a visual analysis of each study to determine the following: (a) the consistency of level, trend, and variability within each phase; (b) how immediate the effect was between baseline and intervention phases, the proportion of overlap, and the consistency of data across phases; and (c) whether “anomalies” existed within the data (Kratochwill et al., 2010). Together, evidence of a functional relation was used to determine whether each study provided *strong evidence*, *moderate evidence*, or *no evidence* of an effect (Kratochwill et al., 2010; Maggin et al., 2011; see Table 1).

Most of the studies were published in peer-reviewed journals; 4 were dissertations. Nine of the studies from the GBG literature reviews were included; 12 studies included in the Flower et al. (2014) meta-analysis were analyzed in the current meta-analysis. Two of the authors independently coded each study using the aforementioned rubric; discrepancies were discussed and resolved. Initial agreement percentages for the application of the four design standards and visual analysis were 88% and 91%, respectively. Final agreement was 100% for both. Agreement for article inclusion was 100%. The formula sum of agreement / total

**Table 1.** Overall Effect of the GBG.

Study	TauU (95% CI)			n	Number of phase contrasts	Number of cases	WWC design standards							Overall <sup>a</sup>
	LL	ES	UL				DS1	DS2	DS3	DS4	Vis			
Parrish (2012)	0.25	0.42	0.60	373	10	1	MS	MSR	MS	MSR	MS	MS	M	MSR
Kosiec, Czernicki, and McLaughlin (1986)	0.43	0.53	0.63	54	16	2	MS	MSR	MS	MS	MS	MS	M	MSR
Medland and Stachnik (1972)	0.55	0.75	0.94	28	2	1	MS	MSR	MS	MS	NM	NM	M	NM
Wright and McCurdy (2011)	0.61	0.89	1.00	37	4	2	MS	MSR	MSR	MSR	MS	MS	M	MSR
Rodriguez (2010)	0.70	0.90	1.00	22	5	5	MS	MSR	MS	MS	MS	MS	M	MSR
Hunt (2012)	0.67	0.93	1.00	57	8	3	MS	MS	MSR	MSR	MS	MS	S	MSR
Salend, Reynolds, and Coyle (1989)	0.84	0.97	1.00	19	14	3	MS	MS	MS	MS	MS	MS	S	MS
Kleinman and Saigh (2011)	0.68	0.97	1.00	26	6	1	MS	MS	MS	MS	MS	MS	M	MS
Saigh and Umar (1983)	0.70	0.98	1.00	20	6	1	MS	MSR	MS	MS	NM	NM	S	NM
Ruiz-Olivares, Pino, and Herruzo (2010)	0.59	0.98	1.00	15	4	1	MS	MS	MSR	MSR	MS	MS	S	MSR
McCurdy, Lannie, and Barnabas (2009)	0.65	0.98	1.00	600	3	3	MS	MS	MS	MS	MS	MS	M	MS
Donaldson, Vollmer, Krouse, Downs, and Berard (2011) <sup>b</sup>	0.78	0.98	1.00	98	5	5	MS	MS	MS	MS	NM	NM	S	NM
Patrick, Ward, and Crouch (1998)	0.76	0.99	1.00	67	6	3	MS	MSR	MSR	MSR	MS	MS	M	MSR
Gresham and Gresham (1982)	0.62	1.00	1.00	12	4	1	MS	MS	MS	MS	MS	MS	S	MS
Tanol, Johnson, McComas, and Cote (2010)	0.73	1.00	1.00	6	8	2	MS	MSR	MSR	MSR	MS	MS	M	MSR
Patterson (2003)	0.84	1.00	1.00	6	6	1	MS	MS	MS	MS	MS	MS	S	MS
Nolan, Filter, and Houlihan (2013)	0.70	1.00	1.00	6	6	3	MS	MSR	MSR	MSR	MS	MS	M	MSR
Crouch, Gresham, and Wright (1985)	0.68	1.00	1.00	22	6	1	MS	MSR	MSR	MSR	NM	NM	M	NM
Bostow and Geiger (1976)	0.52	1.00	1.00	31	2	1	MS	MS	MS	MS	NM	NM	S	NM
Barrish, Saunders, and Wolf (1969)	0.64	1.00	1.00	24	4	1	MS	MSR	MSR	MSR	NM	NM	M	NM
Johnson, Turner, and Konarski (1978)	0.77	1.00	1.00	31	4	2	MS	MS	MS	MS	NM	NM	S	NM
Overall	0.78	0.82	0.87	1,580	137	43								

Note. GBG = Good Behavior Game; CI = confidence interval; LL = lower limit; ES = effect size; UL = upper limit; WWC = What Works Clearinghouse; DS = design standard; DS1 = GBG was systematically manipulated; DS2 = experimental effect across three points in time or three phase changes; DS3 = appropriate number of data points per phase (as described in the text); DS4 = IOA was at least 80% and was reported for at least 20% of baseline and/or intervention phases; Vis = visual analysis; MS = meets standards; MSR = meets standards with reservations; NM = does not meet standards; M = moderate evidence; S = strong evidence; IOA = interobserver agreement.

<sup>a</sup>Includes DS1, DS2, DS3, and DS4. <sup>b</sup>In Donaldson et al. (2011), IOA data met standards for two of the five participating teachers but did not meet standards for three of the teachers.

number of agreements + disagreements  $\times$  100 (House, House, & Campbell, 1981) was used for these and all other instances of IOA.

### Coding of Studies and Intercoder Reliability

The first author operationally defined and coded all 21 studies in an Excel spreadsheet. Two of the co-authors, doctoral students with training in SCR methodology and experience conducting SCR meta-analyses, were trained on the codes. Each student independently coded a set of studies using a separate Excel spreadsheet. Thus, each study was double- or triple-coded. Reliability was calculated for 75% of the studies across 15 study variables including the following: the five potential moderator variables, number of participants, IOA, and fidelity. Initial agreement was 96%. Disagreements were resolved after the first author and doctoral students reread and discussed the articles, resulting in 100% final agreement across all codes. IOA procedures were similar to those reported by Methe, Kilgus, Neiman, and Riley-Tillman (2012).

### Publication Bias and Fixed-Effects Model

Publication bias was statistically tested in WinPepi (Abramson, 2011) using the Egger's test (Egger, Smith, Schneider, & Minder, 1997). The intercept for the Egger's test (2.88, 90% CI = [1.50, 4.16],  $p = .01$ ) suggested publication bias. However, sensitivity analyses conducted in WinPepi indicated that no single study had an undue impact on the findings. Heterogeneity was measured using Higgins' and Thompson's  $H$  and  $I^2$  statistics (Higgins & Thompson, 2002), where  $H = 2.0$  (95% CI = [1.6, 2.5]) and  $I^2 = 75.5\%$  (95% CI = [62.7, 84.0]). While these results indicate evidence of considerable heterogeneity, caution is warranted for two reasons. First, "The [ $H$ ] test has poor power with few studies . . . it can therefore be difficult to decide whether heterogeneity is present or whether it is clinically important" (Higgins & Thompson, 2002, p. 1552). Second, regarding statistical heterogeneity, "there may be situations when the fixed-effects analysis is appropriate even when there is substantial heterogeneity of results (e.g., when the question is specifically about the particular set of studies that have already been conducted)" (Hedges & Vevea, 1998, p. 487).

Neither a fixed-effects nor a random-effects model is an "exact fit" for SCR data, but a fixed-effects model was preferable because the number of cases was too small for reasonable estimation of variances under the random-effects model (Greenhouse & Iyengar, 2009). Thus, the TauU effect size was calculated within a fixed-effects model (see Parker, Vannest, Davis, & Sauber, 2011) using WinPepi. Rather than being regarded as random samples, the studies in this meta-analysis were all regarded as estimates of an unknown "true" effect size. Variations in the "true" effect size were sought through moderator analysis (Hedges & Olkin, 1985).

### Effect Size Estimation

**TauU.** TauU is an effect size measure based on non-overlap between A and B phases. One of its strengths is that it can control for confounding baseline trends (Parker et al., 2011). It performs reasonably well with autocorrelation ( $r_{\text{auto}}$ ); its  $SE$  and significance level are not affected. When tested, 75% of TauU values remained unchanged after  $r_{\text{auto}}$  was cleansed (Parker et al., 2011). TauU is derived from the Kendall's Rank Correlation and the Mann-Whitney  $U$  test between groups. The Kendall's Rank Correlation is an analysis algorithm of time and score, comparing ordered scores and all possible pairs of data. Each pairwise comparison represents an improved score, a score that has not improved, or a tie. The Mann-Whitney  $U$  index represents differences in group level. With regard to SCR, the concept is applied to phases rather than groups, and scores from two phases are combined for a cross-group ranking. The rankings are statistically compared for mean differences. The Mann-Whitney  $U$  algorithm uses two continuous variables: scores and time. Replacing the time variable with a dummy code (0/1) to represent A and B phases yields an identical result. This, in turn, produces the proportion of pairwise comparisons that improve from Phase A to Phase B. TauU is better suited to short phases than most other methods (e.g., techniques relying on linear trends) because it can find reliable monotonic trend in only three or four data points.

**Phase contrasts and effect size calculation.** We used the GetData Digitizer program (version 2.25; <http://www.getdata-graph-digitizer.com/>) to scan and code graphed data. Graphed data from A and B phases were extracted from each study and transformed into raw numerical data by setting a scale based on the  $X$  and  $Y$  values for each phase. Effect size calculation involved several steps. First, an effect size was calculated for each AB contrast (e.g., an effect size for the A1/B1 contrast and a separate effect size for the A2/B2 contrast). Second, digitized data values were entered into the TauU calculator (Vannest, Parker, & Gonan, 2011) to obtain TauU and its standard error ( $SE_{\text{Tau}}$ ). Third, TauU and  $SE_{\text{Tau}}$  values were entered into WinPepi using the meta-analysis function to aggregate the data and arrive at an effect size, standard error, and CI for each study. Fourth, TauU and  $SE_{\text{Tau}}$  values for each study were entered into WinPepi to obtain an omnibus effect size with standard error and CI. Finally, separate TauU,  $SE_{\text{Tau}}$ , and CI values were calculated for each level of each potential moderator (see Table 2).

**TauU phase contrast intercoder agreement.** Each of the 21 studies included multiple phase contrasts (e.g., A1/B1, A2/B2), resulting in 137 phase contrasts. The first author trained the doctoral students in calculating TauU and  $SE_{\text{Tau}}$  for each phase contrast using the TauU calculator from the obtained GetData values. Two of the authors independently

**Table 2.** Effect of the GBG Across Potential Moderators.

Potential moderators	Number of participants	Number of cases	Confidence interval (95%)			z	p
			LL	ES	UL		
Grade level							
Elementary	1,481	37	0.82	0.88	0.94		
Secondary	66	5	0.85	0.97	1.00	1.34	.18
GBG format <sup>a</sup>							
Not modified	321	13	0.74	0.81	0.88		
Modified	1,325	31	0.76	0.82	0.88	0.20	.84
Target behaviors <sup>b</sup>							
Off-task	1,140	39	0.76	0.81	0.86		
On-task	497	7	0.47	0.59	0.72	2.89	.01*
Reinforcer frequency							
Daily	984	30	0.77	0.82	0.88		
Daily and weekly	223	12	0.82	0.92	1.00	1.71	.08
EBD status							
EBD/EBD risk	88	11	0.89	0.98	1.00		
No EBD	1,492	34	0.70	0.76	0.81	3.77	.01*

Note. GBG = Good Behavior Game; LL = lower limit; ES = effect size; UL = upper limit; EBD = emotional and behavioral disorder.

<sup>a</sup>Gresham and Gresham (1982) and Kosiec, Czernicki, and McLaughlin (1986) used original and modified versions of the GBG. The number of participants for these studies are represented in both formats. <sup>b</sup>Hunt (2012) and Patrick, Ward, and Crouch (1998) reported both on- and off-task behavior data. The number of participants from these studies are represented in both types of behavior.

\* $p = .05$

calculated these values for 20% ( $n = 27$ ) of the 137 AB phase contrasts across all studies. Initial agreement for non-overlap between A and B phases ranged from 50% to 100%. The 50% agreement reflects difficulty extracting the data from the Johnson, Turner, and Konarski (1978) study. Disagreements were resolved after the authors discussed the discrepancies and recoded the data, resulting in 100% final agreement. We then calculated TauU and  $SE_{\text{Tau}}$  for the remaining phase contrasts in each study.

**Statistical significance.** We determined statistical significance for TauU values using 95% CI ( $\alpha = .05$ ). A 90% to 95% CI is standard for determining whether change is reliable (Nunnally & Bernstein, 1994), indicating a reasonable chance of 5% to 10% likelihood of error. Statistical significance between TauU values was determined by calculating 83.4% CI to visually test for overlap of upper and lower limits between effect sizes. Visual comparison of two effect sizes with 83.4% CI is the same as a  $p = .05$  or a 95% confidence level test between the two scores (Payton, Greenstone, & Schenker, 2003).

### Potential Moderators

We examined five potential moderators, variables hypothesized to affect students' behaviors. They were selected because they were the recommended areas of future research or had not yet been addressed in the previous reviews or meta-analysis. Potential moderators were as follows: EBD

risk status, reinforcement frequency, target behaviors, GBG format, and grade level.

We calculated a reliable difference for the levels of each potential moderator. If statistically significant differences were obtained between levels, the potential moderator was confirmed as a moderator because it differentially affected students' outcomes. The reliable difference formula,  $(L1 - L2) / \sqrt{[(SE_{\text{Tau}1})^2 + (SE_{\text{Tau}2})^2]}$ , based on the  $t$  test, was used to determine whether levels of a given moderator differed statistically from one another. A reliable difference is one that is so large that it cannot be accounted for solely by chance, given the number of participants and data points. Alpha was set at .05 and the confidence level was set at 95% to determine whether the findings were credible (viz., whether they would change substantially over several retestings). Reliable difference  $z$  test scores and  $p$  values are reported.

**EBD risk status.** The codes used for EBD risk status were *EBD/EBD risk* and *no EBD/no EBD risk*. *EBD/EBD risk* referred to students identified in a given study as having an emotional and/or behavioral disorder, or those at risk for being identified as having an emotional and/or behavioral disorder. Data for students not identified as individuals with or at risk for EBD were coded *no EBD/no EBD risk*.

**Frequency of reinforcement.** In some studies, daily reinforcement was awarded to the winning team(s) at the end of the class period in which the Game was played. In other

studies, it was awarded at the end of the school day if it was implemented in multiple classes. Both daily and weekly reinforcers were awarded to the winning team(s) that met criteria as an extra incentive in some studies. Levels were *daily* and *daily and weekly*.

**Target behaviors.** Target behaviors consisted of two categories: *disruptive/off-task* and *attention to task/on-task*. Disruptive/off-task behaviors included a range of behaviors including the following: being out-of-seat, talking without permission, interrupting, fighting, name-calling, cursing, pushing, hitting, and destroying property. Attention to task/on-task behaviors referred to complying with teacher directions, working quietly, raising one's hand before asking a question, and getting instructional materials without talking.

**GBG format.** GBG format referred to the use of the GBG as originally described by Barrish et al. (1969) or a modification thereof. Levels were *modified* and *not modified*.

**Grade level.** Grade level was represented by two levels: *elementary* (pre-kindergarten through Grade 5) and *secondary* (Grades 6 through 12).

## Results

### Study Characteristics

The 21 SCR studies examined in this meta-analysis consisted of 43 cases and 137 phase contrasts representing 1,580 participants in pre-kindergarten through Grade 12. The majority of the studies focused on elementary school students ( $n = 17$ ), 2 studies targeted secondary students, and 2 studies included elementary and secondary students. Participant gender was reported in 8 studies (341 males, 300 females). Participant ethnicity was reported in 4 studies, representing more than 13 ethnic groups, including African American, Caucasian, Hispanic, Asian, Native American, Mestizo, Creole, and Mayan. Although most of the studies took place in the United States, they were also carried out in Spain (Ruiz-Olivares et al., 2010), Sudan (Saigh & Umar, 1983), British Columbia (Kosiec, Czernicki, & McLaughlin, 1986), and Belize (Nolan, Filter, & Houlihan, 2013). One study (Tanol, Johnson, McComas, & Cote, 2010) conducted in a U.S. school emphasized Native American culture. Participants included students with intellectual disabilities (Gresham & Gresham, 1982), developmental disabilities (Patterson, 2003), EBD (Salend et al., 1989), students at risk for EBD (Tanol et al., 2010), and students without disabilities (Nolan et al., 2013).

Studies were conducted in general education classrooms ( $n = 14$ ), special education classrooms ( $n = 2$ ), school cafeterias ( $n = 2$ ), a Head Start classroom ( $n = 1$ ), a physical

education class ( $n = 1$ ), and a classroom for students experiencing behavioral challenges ( $n = 1$ ). Behavioral rules and expectations were clearly defined in all studies. Content areas and activities during which the GBG was implemented included math, reading, science, social studies, language arts, art, and circle time. Eleven studies implemented a modified GBG format; 8 used the original format. Two studies (Gresham & Gresham, 1982; Kosiec et al., 1986) compared the use of the original format with a modified version. Data for original and modified formats from each of the 2 studies were analyzed separately with separate TauU and  $SE_{\text{TauU}}$  values. All studies reported direct observations of student behaviors. Twenty studies reported IOA (range = 80%–100%). The remaining study (Bostow & Geiger, 1976) reported collecting IOA data, but did not provide them. Only 14 of the 21 studies reported the percentage of observations included in calculating IOA (range = 25%–52% across baseline and/or intervention phases). Nine studies reported fidelity of implementation; average fidelity was 88%. Social validity was reported for teachers and/or students in 13 studies; ratings and interviews indicated high social validity.

### Overall Effect

In response to the first research question, the overall effect of the GBG was examined across the 21 studies, yielding a mean TauU effect size of .82 ( $SE = .02$ , 95% CI = [.78, .87],  $p = .05$ ). To help with effect size interpretation, we transformed the obtained TauU value to a Cohen's  $d$  of 1.99 using the formula  $d = 3.464 \times [1 - \sqrt{1 - \text{TauU}}]$  (Rosenthal, 1994). This would be considered a large effect size based on the commonly accepted values proposed by Cohen (1992). Table 1 presents the range of effect sizes with CIs across studies at a 95% confidence level. Thus, there is a 95% certainty that the true value for the obtained effect size fell between the upper and lower limits of the calculated CI.

### Potential Moderators

We calculated levels of potential moderators using the reliable difference formula. A 95% CI ( $\alpha = .05$ ) was set for each effect size;  $p$  values are reported for  $z$  test values. The results addressed the second research question (see Table 2).

**EBD risk status.** Students with or at risk for EBD yielded a larger effect size ( $ES = .98$ ,  $SE = .05$ , 95% CI = [0.89, 1.00],  $p = .05$ ) than students not identified with or not at risk for EBD ( $ES = .76$ ,  $SE = .03$ , 95% CI = [0.70, 0.81],  $p = .05$ ). Reliable difference values were  $z = 3.77$ ,  $p = .01$ .

**Reinforcement frequency.** The use of both daily and weekly reinforcement resulted in a larger effect size ( $ES = .92$ ,  $SE$

= .05, 95% CI = [0.82, 1.00],  $p = .05$ ) than did the use of daily reinforcement alone for winning teams ( $ES = .82$ ,  $SE = .03$ , 95% CI = [0.77, 0.88],  $p = .05$ ). Reliable difference results for this variable were  $z = 1.71$ ,  $p = .08$ .

**Target behaviors.** A larger effect size was obtained for disruptive/off-task behaviors ( $ES = .81$ ,  $SE = .03$ , 95% CI = [0.76, 0.86],  $p = .05$ ) than for attention to task/on-task behaviors ( $ES = .59$ ,  $SE = .07$ , 95% CI = [0.47, 0.72],  $p = .05$ ). The reliable difference values were  $z = 2.89$ ,  $p = .01$ .

**GBG format.** Interventions using a modified format had a slightly larger effect size ( $ES = .82$ ,  $SE = .03$ , 95% CI = [0.76, 0.88],  $p = .05$ ) than GBG interventions using the original format ( $ES = .81$ ,  $SE = .04$ , 95% CI = [0.74, 0.88],  $p = .05$ ). Reliable difference values were  $z = .20$ ,  $p = .84$ .

**Grade level.** A larger effect size was observed for secondary students ( $ES = .97$ ,  $SE = .06$ , 95% CI = [0.85, 1.00],  $p = .05$ ) than for elementary students ( $ES = .88$ ,  $SE = .03$ , 95% CI = [0.82, 0.94],  $p = .05$ ). Kosiec et al. (1986) included both elementary and secondary participants. However, their study was not included in the grade-level analysis because the data were not disaggregated for elementary and secondary students. The reliable difference  $z$  test value for grade level was 1.34,  $p = .18$ .

## Discussion

The purpose of this meta-analysis was to examine the effect of the GBG and five potential moderators on elementary and secondary students' behaviors across 21 SCR studies. Specifically, this is the first meta-analysis of the GBG to examine (a) disability (viz., EBD), (b) GBG format, and (c) reinforcement frequency as potential moderator variables. There were several findings worth noting. First, the large overall effect ( $ES = .82$ ) indicated that a reduction in problem behaviors and an increase in desirable behaviors may be attributed to the GBG. Second, moderator analyses revealed a statistically significant difference for two variables: EBD risk status and target behaviors. That is, students with or at risk for EBD benefited more from the GBG than their peers without EBD. Also, students who exhibited disruptive and off-task behaviors benefited most from the Game. Although there were more participants in the disruptive/off-task behaviors group (see Table 2), TauU weighs the number of observations in each study by the inverse of the variance. As such, findings were not influenced by moderator group *ns*. Third, findings also revealed that the GBG was more effective in reducing disruptive/off-task behaviors than increasing attention to task/on-task behaviors.

Although the remaining three potential moderators were not statistically significant, our results were similar to the findings reported by Flower et al. (2014) for grade level. We found moderate to large effects of the GBG in reducing

problem behaviors for elementary and secondary students. We also discovered that, consistent with conclusions reported by Tankersley (1995) and Tingstrom et al. (2006), the GBG implemented in both its original and modified formats was effective. This offers teachers some flexibility in tailoring the Game to their students' behavioral needs. For example, the Game can be modified to award points for appropriate behaviors rather than deduct points for inappropriate behaviors (Crouch et al., 1985), use three teams versus two (Hunt, 2012), implement the intervention in a non-classroom setting (e.g., the cafeteria; McCurdy, Lannie, & Barnabas, 2009), or include an additional behavioral intervention (Ruiz-Olivares et al., 2010). However, we caution that although there seems to be some flexibility in its implementation, the core features of the Game should be adhered to as an interdependent group contingency to achieve outcomes such as those noted in the literature.

With regard to reinforcement frequency, there was a greater reduction in problem behaviors with more frequent reinforcement. This finding is particularly noteworthy for students with or at risk for EBD (Cheyney & Jewell, 2012). Last, like Flower et al., we noted that very few studies reported fidelity of implementation. It is important to report these data to help understand the degree to which and the consistency with which the GBG is implemented. Such data could inform revisions or improve implementation of the intervention.

## Limitations

The following limitations should be considered in interpreting the findings of this meta-analysis. First, there are no generally agreed upon standards within SCR on the use of meta-analysis within the field to determine evidence-based practices (Horner & Kratochwill, 2012). Although the use of meta-analysis in other fields to synthesize research literature is becoming a standard practice (Parker & Hagan-Burke, 2007), its application in applied behavior analysis and SCR is more limited. Second, we did not use IOA as an inclusion criterion because so few studies reported the information needed to apply this proposed design standard. As new criteria are developed and existing protocols are revised, studies may be judged differently by others in the field. Third, effects on individual students could not be determined in some studies. For example, Barrish et al. (1969) reported that the two most behaviorally challenging students received marks for inappropriate behaviors individually rather than deducting points from their team for their disruptive behaviors and refusal to participate in the Game. Thus, the impact of the Game on all students is not reflected in the data reported in this study. Fourth, most studies did not disaggregate behaviors by type (e.g., aggressive). Rather, a range of problem behaviors were operationally defined and combined in a "disruptive behavior" category. As such, results could not be provided for some categories of problem behaviors. Finally, as a newer effect

size measure, TauU has been used in few meta-analyses (e.g., Bowman-Perrott et al., 2013) and intervention studies (e.g., Rispoli et al., 2013). Also, caution should be used in comparing TauU with Cohen's  $d$ , as the transformation is an approximation.

### Implications for Research and Practice

As additional SCR studies are conducted, complete IOA data (viz., the percentage of observations included in IOA calculations) should be reported. In addition, future research should investigate the potential mediating effect of several variables. One is gender, as boys have been found to be more likely to display disruptive behaviors than girls (McIntosh, Reinke, Kelm, & Sadler, 2013). A second variable is ethnicity, as it is related to other types of behavioral outcomes (e.g., suspension and expulsion from school; Bowman-Perrott et al., 2011). Intervention length and duration may also make a difference in students' outcomes (Kosiec et al., 1986). In future GBG studies, behavioral outcome data need to be disaggregated by type, as several studies combined physically and verbally aggressive behaviors with out-of-seat and talking out behaviors. In addition, the impact of the GBG on students' academic achievement (Flower et al., 2014; Kellam et al., 1994) should be examined in consideration of the relation between academic difficulties and behavior problems (Lane, Barton-Arwood, Nelson, & Wehby, 2008). Finally, the use of response cost versus reinforcement in promoting positive behaviors during the GBG (Tanol et al., 2010; Wright & McCurdy, 2011) should be further examined.

Identifying empirically supported practices is important in an era of increased accountability. It is critical that school personnel identify and implement classroom management and behavioral interventions that promote prosocial behaviors. In light of the evidence pointing to the GBG as an effective universal behavior management strategy, it can be a benefit to general and special education teachers. Overall, the GBG is an effective, positive behavioral support that can easily be incorporated into elementary and secondary school-based settings.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### References

\*References with an asterisk indicate articles included in the meta-analysis.

- Abramson, J. H. (2011). WINPEPI updated: Computer programs for epidemiologists and their teaching potential. *Epidemiologic Perspectives & Innovations*, 8(1), 1–9.
- Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, 31, 621–631.
- Babiyak, A. E., Luze, G. J., & Kamps, D. M. (2000). The Good Behavior Game: Behavior management for diverse classrooms. *Intervention in School and Clinic*, 35, 216–233.
- \*Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good Behavior Game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis*, 2, 119–124.
- \*Bostow, D., & Geiger, O. G. (1976). Good Behavior Game: A replication and systematic analysis with a second grade class. *School Applications of Learning Theory*, 8(2), 18–27.
- Bowman-Perrott, L., Benz, M. R., Hsu, H., Kwok, O., Eisterhold, L., & Zhang, D. (2011). Patterns and predictors of disciplinary exclusion over time: An analysis of the SEELS national dataset. *Journal of Emotional and Behavioral Disorders*, 21(2), 83–96.
- Bowman-Perrott, L., Davis, H., Vannest, K. J., Williams, L., Greenwood, C. R., & Parker, R. (2013). Academic benefits of peer tutoring: A meta-analytic review of single-case research. *School Psychology Review*, 42(1), 39–55.
- Cheyney, D., & Jewell, K. (2012). Positive behavior supports and students with emotional and behavioral disorders. In J. P. Bakken, F. E. Obiakor, & A. F. Rotatori (Eds.), *Behavioral disorders: Practice concerns and students with EBD* (Vol. 23; pp. 83–106). Bingley, UK: Emerald.
- Cohen, J. (1992). A power primer. *Quantitative Methods in Psychology*, 112, 155–159.
- Cooper, H. (2011). *Reporting research in psychology: How to meet journal article reporting standards* (6th ed.). Baltimore, MD: American Psychological Association.
- \*Crouch, P. L., Gresham, F. M., & Wright, W. R. (1985). Interdependent and independent group contingencies with immediate and delayed reinforcement for controlling classroom behavior. *Journal of School Psychology*, 23, 177–187.
- Darveau, D. X. (1984). The Good Behavior Game plus merit: Controlling disruptive behavior and improving student motivation. *School Psychology Review*, 13, 510–514.
- Dolan, L. J., Kellam, S. G., Brown, C. H., Werthamer-Larsson, L., Rebok, G. W., . . . Wheeler, L. (1993). The short-term impact of two classroom-based preventative interventions on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology*, 14, 317–345.
- \*Donaldson, J. M., Vollmer, T. R., Krouse, T., Downs, S., & Berard, K. P. (2011). An evaluation of the Good Behavior Game in kindergarten classrooms. *Journal of Applied Behavior Analysis*, 44, 605–609.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634.
- Embry, D. D. (2002). The Good Behavior Game: A best practice candidate as a universal behavioral vaccine. *Clinical Child and Family Psychology Review*, 5, 273–297.
- Fishbein, J. E., & Wasik, B. H. (1981). Effect of the Good Behavior Game on disruptive library behavior. *Journal of Applied Behavior Analysis*, 1(14), 89–93.

- Flower, A., McKenna, J. W., Bunuan, R. L., Muething, C. S., & Vega, R. (2014). Effects of the Good Behavior Game on challenging behaviors in school settings. *Review of Educational Research, 84*, 546–571. doi:10.3102/0034654314536781
- Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 417–433). New York, NY: Russell Sage.
- \*Gresham, F. M., & Gresham, G. N. (1982). Interdependent, dependent, and independent group contingencies for controlling disruptive behavior. *The Journal of Special Education, 16*, 101–110.
- Harris, V. W., & Sherman, J. A. (1973). Use and analysis of the “Good Behavior Game” to reduce disruptive classroom behavior. *Journal of Applied Behavior Analysis, 6*, 405–417.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486–504.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539–1558.
- Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: Some brief reflections. *Journal of Behavioral Education, 21*, 266–272.
- House, A. E., House, B. J., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Journal of Behavioral Assessment, 3*(1), 37–57.
- Huber, H. (1979). The value of a behavior modification program administered in the fourth grade of a school for special education. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 28*(2), 73–79.
- Huizink, A. C., van Lier, P. A. C., & Crijnen, A. A. M. (2008). Attention deficit hyperactivity disorder symptoms mediate early-onset smoking. *European Addiction Research, 15*, 1–9.
- \*Hunt, B. M. (2012). Using the Good Behavior Game to decrease disruptive behavior while increasing academic engagement with a Head Start population (Unpublished doctoral dissertation). University of Southern Mississippi, Hattiesburg.
- \*Johnson, M. R., Turner, P. F., & Konarski, E. A. (1978). The “Good Behavior Game”: A systematic replication in two unruly transitional classrooms. *Education and Treatment of Children, 1*(3), 25–33.
- Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, 35*, 259–281.
- \*Kleinman, K. E., & Saigh, P. A. (2011). The effects of the Good Behavior Game on the conduct of regular education New York city high school students. *Behavior Modification, 35*, 95–105.
- \*Kosiec, L. E., Czernicki, M. R., & McLaughlin, T. F. (1986). The Good Behavior Game: A replication with consumer satisfaction in two regular elementary school classrooms. *Techniques, 2*(1), 15–23.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from [http://ies.ed.gov/ncee/wwc/pdf/wwc\\_scd.pdf](http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf)
- Lane, K. L., Barton-Arwood, S., Nelson, R., & Wehby, J. (2008). Academic performance of students with emotional and behavioral disorders in a self-contained setting. *Journal of Behavioral Education, 17*, 43–62.
- Leflot, G., van Lier, P. A. C., Onghena, P., & Colpin, H. (2010). The role of teacher behavior management in the development of disruptive behaviors: An intervention study with the Good Behavior Game. *Journal of Abnormal Child Psychology, 38*, 869–882.
- Maggin, D. M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. J. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology, 49*, 529–544.
- \*McCurdy, B. L., Lannie, A. L., & Barnabas, E. (2009). Reducing disruptive behavior in an urban school cafeteria: An extension of the Good Behavior Game. *Journal of School Psychology, 47*(1), 39–54.
- McGoey, K. E., Schneider, D. L., Rezzetano, K. M., Prodan, T., & Tankersley, M. (2010). Classwide intervention to manage disruptive behavior in the kindergarten classroom. *Journal of Applied School Psychology, 26*, 247–261.
- McIntosh, K., Reinke, W. M., Kelm, J. L., & Sadler, C. A. (2013). Gender differences in reading skill and problem behavior in elementary school. *Journal of Positive Behavior Interventions, 15*, 51–60.
- \*Medland, M. B., & Stachnik, T. J. (1972). Good-Behavior Game: A replication and systematic analysis. *Journal of Applied Behavior Analysis, 5*(1), 45–51.
- Methe, S. A., Kilgus, S. P., Neiman, C., & Riley-Tillman, T. C. (2012). Meta-analysis of interventions for basic mathematics computation in single-case research. *Journal of Behavioral Education, 21*, 230–253.
- \*Nolan, J. D., Filter, K. J., & Houlihan, D. (2013). Preliminary report: An application of the Good Behavior Game in the developing nation of Belize. *School Psychology International, 35*, 421–428. doi:10.1177/014303498598
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw Hill.
- Parker, R., & Hagan-Burke, S. (2007). Useful effect sizes interpretations for single case research. *Behavioral Therapy, 38*, 95–105.
- Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining non-overlap and trend for single-case research: Tau-U. *Behavior Therapy, 35*, 202–322.
- \*Parrish, R. (2012). Examining changes in appropriate social behaviors during school lunch using the lunchtime behavior game (Unpublished doctoral dissertation). Northeastern University, Boston, MA.
- \*Patrick, C. A., Ward, P., & Crouch, D. W. (1998). Effects of holding students accountable for social behaviors during volleyball games in elementary physical education. *Journal of Teaching in Physical Education, 17*, 143–156.
- \*Patterson, K. B. (2003). The effects of a group-oriented contingency—the Good Behavior Game—on the disruptive behavior of children with developmental disabilities (Unpublished doctoral dissertation). Kent State University, Kent, OH.

- Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science*, 3(4), 1–6.
- Reid, R., Gonzalez, J. E., Nordess, P. D., Trout, A., & Epstein, M. (2004). A meta-analysis of the academic status of students with emotional/behavioral disturbance. *The Journal of Special Education*, 38, 130–143.
- Rispoli, M., Lang, R., Neely, L., Camargo, S., Hutchins, N., Davenport, K., & Goodwyn, F. (2013). A comparison of within- and across-activity choices for reducing challenging behavior in children with autism spectrum disorders. *Journal of Behavioral Education*, 22, 66–83.
- \*Rodriguez, B. J. (2010). *An evaluation of the Good Behavior Game in early reading intervention groups* (Unpublished doctoral dissertation). University of Oregon, Eugene.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 232–243). New York, NY: Russell Sage.
- Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52, 59–82.
- \*Ruiz-Olivares, R., Pino, M. J., & Herruzo, J. (2010). Reduction of disruptive behaviors using an intervention based on the good behavior game and the say-do-report correspondence. *Psychology in the Schools*, 47, 1046–1058.
- \*Saigh, P. A., & Umar, A. M. (1983). The effects of a Good Behavior Game on the disruptive behavior of Sudanese elementary school students. *Journal of Applied Behavior Analysis*, 16, 339–344.
- \*Salend, S. J., Reynolds, C. J., & Coyle, E. M. (1989). Individualizing the Good Behavior Game across type and frequency of behavior with emotionally disturbed adolescents. *Behavior Modification*, 13, 108–126.
- Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 2(3), 188–196.
- Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children*, 31, 351–380.
- Swiezy, N. B., Matson, J. L., & Box, P. (1992). The Good Behavior Game: A token reinforcement system for preschoolers. *Child & Family Behavior Therapy*, 14(3), 21–32.
- Tankersley, M. (1995). A group-oriented contingency management program: A review of research on the Good Behavior Game and implications for teachers. *Preventing School Failure*, 40, 19–24.
- \*Tanol, G., Johnson, L., McComas, J., & Cote, E. (2010). Responding to rule violations or rule following: A comparison of two versions of the Good Behavior Game with kindergarten students. *Journal of School Psychology*, 48, 337–355.
- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44, 423–432.
- Tingstrom, D. H., Sterling-Turner, H. E., & Wilczynski, S. M. (2006). The Good Behavior Game: 1969-2002. *Behavior Modification*, 30, 225–253.
- Vannest, K. J., Parker, R. I., & Gonan, O. (2011). Single-case research: Web-based calculators for SCR analysis, Version 1.0 [Web-based application]. College Station: Texas A&M University. Available from [www.singlecaseresearch.org](http://www.singlecaseresearch.org)
- \*Wright, R. A., & McCurdy, B. L. (2011). Class-wide positive behavior support and group contingencies: Examining a positive variation of the good behavior game. *Journal of Positive Behavior Interventions*, 14, 173–180.