Review of Educational Research
Month 201X, Vol. XX, No. X, pp. 1–26
DOI: 10.3102/0034654314536781

Effects of the Good Behavior Game on Challenging Behaviors in School Settings

Andrea Flower
The University of Texas at Austin

John W. McKenna
St. John's University

Rommel L. Bunuan
Colin S. Muething
Ramon Vega Jr.
The University of Texas at Austin

Challenging behavior at school remains a concern for teachers and administrators. Thus classroom management practices to prevent challenging behavior are sorely needed. The Good Behavior Game (GBG) has been found to be useful to positively change student behavior. However, previous reviews of the GBG have not quantified effects, have focused solely on school and classroom behaviors, and have not examined study features that facilitate greater outcomes. Twenty-two peer-reviewed journal articles were reviewed. Study data were analyzed using effect sizes, percent of nonoverlapping data, percent of all nonoverlapping data, and hierarchical linear modeling to determine intervention effectiveness as well as study features that facilitated greater outcomes. Findings suggested that (a) moderate to large effects were found on challenging behaviors and these effects were immediate; (b) the GBG was most commonly used for disruptive behavior, off-task behavior, aggression, talking out, and out-of-seat behaviors; (c) the GBG has been implemented primarily in general education elementary school settings; and (d) correct application of reward procedures are important for intervention effectiveness. Study limitations, implications for practice, and areas for future research are presented.

Keywords: review, Good Behavior Game, challenging behavior, classroom management, hierarchical linear modeling, effect size

Today, in the context of high-stakes assessments, school reform, and improved academic achievement, teachers are under additional pressure to maximize instructional time to promote academic development (Vannest, Temple-Harvey, & Mason, 2009). Yet, in today's schools, students are not always academically or

socially prepared for school (Blair & Diamond, 2008). In fact, challenging behavior often interrupts teachers' abilities to teach and students' abilities to learn (Cameron, Connor, Morrison, & Jewkes, 2008). With the move toward inclusion and legislation requiring student placement in the least restrictive environment (Individuals with Disabilities Education Improvement Act, 2004), students with challenging behavior might be served in a variety of settings including general education or special education classrooms. This change means that all teachers should be prepared to provide academic instruction as well as behavioral support (Bohannon & Wu, 2011; Witt, VanDerHeyden, & Gilbertson, 2004).

Challenging behavior at school may manifest under many conditions and in various locations throughout a school. A variety of behaviors have been identified by researchers as challenging at school including physical and verbal aggression, harassment, fighting, disrespect, and defiance (Kaufman et al., 2010; Spaulding et al., 2010), getting out of one's seat, talking without permission, and classroom rule violations (Walter, Gouze, & Lim, 2006). Harrison, Vannest, Davis, and Reynolds (2012) found that teachers identified general distractibility and difficulty following directions as the most challenging behaviors. Considering the range of problem behaviors and the time spent handling them, strategies are needed to prevent the occurrence of challenging behavior and promote the display of appropriate behaviors.

The Good Behavior Game

The Good Behavior Game (GBG) is a classroom management strategy that has been used and studied for more than 40 years. The GBG is an easy to implement group contingency procedure that includes identifying target behaviors, posting rules, identifying rewards, dividing a class into at least two equal teams, identifying rule violators and stating their infractions, debiting the offending team for infractions or awarding points for meeting expectations, and awarding daily and weekly prizes to the team with the fewest infractions (Barrish, Saunders, & Wolf, 1969; Elswick & Casey, 2012) or most points earned for prosocial behavior. The GBG allows teachers to engage in several behavior management strategies including acknowledging appropriate behavior, teaching classroom rules, providing feedback about inappropriate behavior, engaging in response cost practices, verbal praise, and providing rewards as reinforcement. Thus, the GBG is a potentially effective classroom management tool for teacher use (Elswick & Casey, 2012).

In its initial empirical evaluation (Barrish et al., 1969), researchers used the GBG to decrease out-of-seat and talking-out behaviors of fourth-grade students during mathematics and reading instruction. Since that initial investigation, the GBG has been applied numerous times to test its effects on a variety of behaviors. The GBG has been implemented by various intervention agents with varying levels of GBG training for different lengths of time, in a variety of settings, and with and without using rewards (cf., Darch & Thorpe, 1977; Elswick & Casey, 2012; McCurdy, Lannie, & Barnabas, 2009; Tanol, Johnson, & McComas, 2010). These factors might make a difference in outcomes as well.

Two earlier reviews (Embry, 2002; Tingstrom, Sterling-Turner, & Wilczynski, 2006) highlighted studies of the GBG, with both reviews concluding that the GBG promotes behavior change. Embry (2002) provided descriptive information

on previously published studies as well as key findings such as intervention effectiveness in terms of immediate behavior change and long-term impact, social validity, and potential use as a prevention strategy. Tingstrom et al. (2006) also wrote a descriptive review in which they discussed the GBG variations used in experimental studies as well as the student populations included in such studies. They also suggested areas for future research such as investigating methods for limiting the amount of peer pressure placed on students who violate game rules and its potential effect on intervention effectiveness.

Neither Embry (2002) nor Tingstrom et al. (2006) specifically focused on observable and measurable challenging behaviors in school or classroom settings. A focus on such behaviors in schools and classrooms is important given the amount of time teachers and administrators spend handling challenging behavior (U.S. Department of Education, 2000), the amount of instructional time lost to these behaviors, and the amount of stress that teachers feel due to challenging behavior (Klassen & Chiu, 2010; Nelson, Maculan, Roberts, & Ohlund, 2001). Additionally, neither Embry nor Tingstrom et al. quantified the effect of the GBG. In our review, we attempt to quantify the effect of the GBG, particularly on observable and measurable challenging behaviors in school or classroom settings. Some of the studies included in our review use group design and others utilize single subject experimental designs (SSEDs).

For group designs, Cohen's *d* is a widely accepted metric for quantifying magnitude of effect. Unfortunately, a method for quantifying effects across SSEDs in meta-analyses is less clear. One promising approach for meta-analysis of SSEDs includes the use of hierarchical linear modeling (HLM; Nagler, Rindskopf, & Shadish, 2008). The HLM approach allows for the aggregation of multiple studies at the case level, which increases overall sample size. Additionally, HLM corrects for autocorrelation and allows for analysis of data from multiple cases, even when the numbers of observations vary across cases (Raudenbush & Bryk, 2002). Using HLM we can study the treatment effect using all of the data from each case.

In a resource-limited system such as a school, knowledge of the expected effects is very desirable before investing the time and resources in any intervention. School professionals might also benefit from information concerning differences in effect of the GBG based on variations in fidelity, intervention agent, duration, setting, and use of rewards. Because HLM analysis does not provide an individual effect size for each study, nonparametric statistics such as percentage of nonoverlapping data (PND) and percentage of all nonoverlapping data (PAND) were useful statistics as we examined how various study characteristics may have affected GBG outcomes.

Purpose and Research Questions

With the importance of student achievement in school, it is essential that teachers manage and change challenging classroom and school behavior so that more time can be allocated to academic instruction. Previous reviews have not quantified the effects of the GBG, nor have they examined specific characteristics of GBG interventions (i.e., fidelity, intervention agent, setting, duration, and/or use of rewards) that may affect outcomes (Embry, 2002; Tingstrom et al., 2006). Additionally, it has been 7 years since researchers have reviewed this literature. In

our review, we attempted to identify additional and more recent experimental studies that examined the effect of the GBG on classroom and school challenging behaviors. The purposes of this review were to (a) describe and quantify the effect of the GBG on various challenging behaviors in school and classroom settings and (b) understand characteristics of the intervention that may affect the magnitude of the outcomes. The following questions guided this review:

Research Question 1: What is the effect of the GBG on the level and trend of challenging behaviors in school and classroom settings?

Research Question 2: Do variations in fidelity, intervention agent, duration, setting, and use of rewards affect GBG outcomes?

Method

We searched the literature to identify studies of the GBG implemented with students in kindergarten through 12th grade between 1970 and the present. Barrish et al. (1969) developed the GBG in 1969; thus, years prior to 1970 were not searched. First, we conducted an electronic search using EBSCO Research Databases including Academic Search Complete, ERIC, and PsycInfo. Next, in an effort to capture all articles concerning the GBG, the reference lists of all GBG articles including the two previous reviews (Embry, 2002; Tingstrom et al., 2006) were searched. In addition, we manually searched several journals that appeared to frequently publish GBG articles in the event that the electronic search failed to identify all relevant articles. The journals included in the manual search were *Behavior Modification*, *Education & Treatment of Children*, *Journal of Applied Behavior Analysis*, *Journal of School Psychology*, and *Psychology in Schools*.

Our initial search yielded 51 articles, including the original empirical investigation of the GBG, which potentially met our inclusion criteria. We then applied a set of inclusion criteria to identify our final pool of articles. Articles selected for inclusion met the following criteria:

1. Published in a peer-reviewed journal in education, special education, behavioral analysis, psychology, or school psychology between 1970 and spring 2013;
2. Article written in English;
3. Referred to the independent variable as the GBG and used GBG procedures;
4. The research design was either an experimental/quasi-experimental design or SSED with replication (multiple baseline or reversal);
5. Dependent variables were challenging behaviors that were a threat to learning, safety, and relationships that were observable and measurable;
6. Results of the GBG implementation could be disaggregated; and
7. Data were available for extraction or calculation of effect size.

Article Coding

Two researchers read and independently double-coded each article with regard to design, dependent variables of the study, outcomes, fidelity, interventionist, interventionist training, duration, setting, and reward use. Design referred to the

research design used in the study. Dependent variables (DVs) were the specific outcome variables that concerned challenging behaviors that were a threat to learning, safety, and relationships, which were observable and measurable in the classroom or school setting. We reviewed the DVs and operational definitions given in each article. Operational definitions were closely matched across studies. For example, talking out might have been referred to as either talking out or inappropriate verbalizations (cf., Barrish et al., 1969; Salend, Reynolds, & Coyle, 1989). Operationally, both behaviors were defined as verbalizations without teacher permission. DVs defined so similarly were categorized as the same DV. Outcomes referred to the results for each DV.

Fidelity referred to the treatment integrity of the intervention, essentially whether GBG took place as intended. GBG interventionist referred to the personnel responsible for implementing the GBG with children. Training for GBG interventionists concerned the amount of training and type of training received prior to and during GBG implementation. Training for the interventionist was coded as lecture, lecture/feedback, modeling, or no response. Lecture was defined as receipt of information about the GBG without additional exposure. Lecture/feedback was defined as receipt of information with feedback from the experimenter after attempting implementation. Modeling was defined as the experimenter demonstrating how to use the GBG in the interventionist's setting.

Duration referred to the number of days for which a study was conducted. When an article reported another duration metric such as number of weeks or months, conversions were made to days in order to standardize the duration code (1 week = 5 school days). Conversions were considered to be estimates, as the authors did not specifically provide them. Four studies (Dion et al., 2011; Leflot, van Lier, Onghena, & Colpin, 2010; Leflot, van Lier, Onghena, & Colpin, 2013; Ruiz-Olivares, Pino, & Herruzo, 2010) required such estimates. Setting pertained to the level of school (elementary or secondary), grade level, type of school or class (e.g., general education, special education, traditional school campus, alternative learning center, school within a residential facility), and location of GBG implementation in the school (e.g., classroom, cafeteria). Finally, reward use concerned use of points (and/or fouls) with teams of students, use of rewards that could be exchanged for points, the type of backup rewards used, frequency of rewards, and use of preference assessments for rewards.

Codes for each component were reviewed and compared for similarity across coders. Calculation of overall and point-by-point reliability relied on this formula (number of agreements divided by agreements plus disagreements, multiplied by 100) to arrive at a percent reliability (Kazdin, 2011). After initial, independent coding, the mean agreement level was 97.2% and point-by-point agreement ranged from 94.3% to 100%. When disagreements occurred, both coders returned to the original article and recoded the source of the disagreement. The coders discussed disagreements and arrived at 100% agreement on each code. Final overall reliability and point-by-point reliability were 100%.

Data Extraction

Two members of the research team extracted data from each graph associated with the 16 SSED studies. The team used GraphClick, a data extraction program,

to extract data from the graphs. Recent research (Boyle, Samaha, Rodewald, & Hoffman, 2012) has validated GraphClick as a method that yields reliable and valid data. From the 16 articles, using GraphClick, we extracted 1,439 data points that corresponded to baseline (582) and GBG (857) intervention conditions.

Use of HLM requires a single dimension across the DVs. For these 16 SSEDs, five studies used frequency as a DV, nine relied on percentage, and two used rate as dimensions for the associated analyses. Analysis with HLM required scaling of the DVs to percentages. Frequency and rate data were converted to percentages by dividing the score represented in each datapoint by the total number possible. Data were also recoded to standardize the direction of intervention effect as some studies aimed to decrease inappropriate behaviors and others aimed to increase appropriate behaviors. Reverse coding of data allowed us to reflect decreases in inappropriate behaviors across all studies. We determined the recoded value by subtracting the old value from the sum of the scale minimum and scale maximum.

Data Analysis

Data Analysis for Research Question 1

Effect size calculations allowed for a determination of the effect of the GBG on challenging school and classroom behavior. We either (a) used Cohen's *d* as provided in the article or (b) calculated the effect size for each group study. Cohen's *d* was calculated as the difference between the mean posttest score of the treatment group minus the mean posttest score of the control group divided by the pooled standard deviation (Cooper, Hedges, & Valentine, 2009). Effect sizes explain the degree to which outcomes for GBG participants differed compared with control group participants. Effect sizes were interpreted using the following criteria: $d = .80$ or greater (large), $d = .50$ (moderate), and $d = .20$ (modest; Cohen, 1988). Averaging across the individual effect sizes provided an overall measure of magnitude.

For SSEDs, we used HLM analysis to determine the effect of the GBG on the level of challenging behavior as well as trend over time. HLM analyses were conducted using SAS PROC MIXED (Little, Milliken, Stroup, & Wolfinger, 1996) to estimate the parameters of interest—β_{2jk} for the treatment effect on the time trend. The following three-level regression model has been suggested (Van den Noortgate & Onghena, 2003) as a method of summarizing such results:

$$\Upsilon_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + \beta_{2jk} T_{ijk} + \beta_{3jk} D_{ijk} T_{ijk} + e_{ijk}.$$

In this equation, the variables represent the following: Υ_{ijk} is the outcome score at measurement occasion i, for subject j, from study k; T_{ijk} is a time-related variable that equals 0 on the first day of the treatment phase; D_{ijk} is a dummy variable that equals 0 in the baseline phase and 1 in the treatment phase; $T_{ijk}D_{ijk}$ is the interaction between T_{ijk} and D_{ijk}; β_{0jk} is the baseline intercept (i.e., the overall mean of the outcome); β_{1jk} is the linear trend during the baseline; β_{2jk} is the treatment effect on the intercept for the trend during the intervention phase (i.e., immediate treatment effect); and β_{3jk} is the treatment effect on the time trend.

Data Analysis for Research Question 2

To answer the second research question concerning whether variations in intervention characteristics (i.e., fidelity, intervention agent, duration, setting, and/or use of rewards) affected GBG outcomes, we examined effect sizes for group designs and calculated and examined PND (Scruggs, Mastropieri, & Castro, 1987) and PAND (Parker, Hagan-Burke, & Vannest, 2007) for all SSEDs. These results allowed us to review whether variations in these variables appeared to affect the outcome data. Data on these study features were compared across studies with modest effects compared with those with moderate or large effects. We were not able to enter the intervention characteristics into the HLM models for analysis because not all data were available for each study.

PND and PAND are nonparametric estimators of effect size commonly used to analyze of SSED studies (Alresheed, Hott, & Bano, 2013). According to the metrics of PND and PAND, little or no overlap between baseline and intervention is considered evidence of a treatment effect (Kratochwill et al., 2002). Calculation of PNDs required counting the number of treatment data points that were greater than the highest data point in baseline, dividing this value by the total number of treatment points, and multiplying this number by 100. For studies in which the expectation was to decrease behavior using the GBG, PND calculations required use of the lowest baseline data point and use of treatment data points below the lowest baseline data point. Criteria for PND effectiveness are as follows: PND less than 50% is considered ineffective, PND between 50% and 70% is considered mildly effective, PND between 70% and 90% is considered moderately effective, and PND greater than 90% is considered highly effective (Mastropieri, Scruggs, Bakken, & Whedon, 1996).

However, PND has some limitations, including the omission of the majority of baseline data points and the overreliance on a single data point that may be an outlier (Parker et al., 2007). PAND serves as a complementary measure of intervention effectiveness because it directly addresses the criticism leveled against PND by using all data points in the analysis. PAND refers to the percent of all data remaining after removing the overlap between the baseline and intervention phases (Parker et al., 2007). The PAND calculation requires identifying the overlapping data points, dividing the number of overlapping data points by the total number of data points and subtracting this number from 100. However, the PAND calculation requires 20 or more data points to have a minimum of five data points for each cell of a 2×2 table, which is the same as used for a chi-square analysis (Parker et al., 2007).

Most of the study features were inspected for whether their qualitative components affected effect size, PND, or PAND. For duration, we calculated and compared the median duration of GBG implementation across studies that indicated a moderate or high effect compared with studies with a null, questionable, or modest effect. Determination of median required ordering the number of days of duration and finding the midpoint for each group.

Results

The purpose of the present review was to (a) describe the strength of effects of the GBG on challenging behaviors in school and classroom settings and (b) to critically examine the differences in outcomes with respect to the intervention

agent, setting, duration of the GBG, reward procedures. In the following sections, we present our findings concerning the effect of the GBG on school and classroom behaviors as well as how characteristics of the intervention have affected the results of each study. First, we review the corpus of studies and the findings of each study according to the challenging behavior on which the researchers focused. Then, we present our findings for each of the two research questions.

Corpus of Studies and Summary of Individual Study Findings

Twenty-two articles published in 14 journals met all inclusion criteria. Articles about the GBG and challenging behaviors in school and classroom settings were most frequently published in the *Journal of Applied Behavioral Analysis* ($n = 5$) and the *Journal of School Psychology* ($n = 3$). Two articles were published in *Psychology in the Schools* and two were published in *Behavior Modification*. The 22 articles included 16 studies with SSEDs and 6 with experimental designs. Four articles reported on two longitudinal studies (Kellam, Ling, Merisca, Brown & Ialongo, 1998; Kellam, Rebok, Ialongo, & Mayer, 1994; Leflot et al., 2010; Leflot et al., 2013).

School and classroom challenging behavior outcomes that were addressed in these articles were disruptive behavior ($n = 8$), off-task/on-task behavior ($n = 6$), aggression ($n = 5$), talking out ($n = 4$), out-of-seat behavior ($n = 4$), peer acceptance and rejection ($n = 2$), rule violations ($n = 2$), antisocial negative behaviors ($n = 1$), appropriate and inappropriate social interactions ($n = 1$), externalizing behavior ($n = 1$), and swearing or negative comments ($n = 1$). Some articles addressed more than one DV with the GBG.

Disruptive Behavior

Disruptive behavior was a combined variable consisting of multiple challenging behaviors such as talking out, out of seat, and touching others or behavior that disrupts activities of another student such as motor activities, noisemaking, verbalizations, or aggression. All the studies that addressed disruptive behavior as a DV used SSEDs to study GBG effects. Across the eight studies (see Table 1) that used the GBG in an effort to reduce disruptive behavior, only Fishbein and Wasik (1981) and Lannie and McCurdy (2007) found the GBG to be ineffective for at least one case in each of their studies. Although McCurdy et al. (2009) indicated that one case experienced an ineffective intervention according to PND, use of PAND for analysis reflected a highly effective intervention.

On-Task and Off-Task

On-task behavior was defined as engaged in tasks as requested, paying attention to academic activities, and visibly engaged in tasks. Off-task referred to behaviors that were incompatible with being engaged in assignments or instruction, failing to pay attention to academic activities, and being visibly unengaged in instructional tasks. On- and off-task behavior were typically mutually exclusive DVs in these studies—that is, a student could not be categorized as on and off-task at the same time. Six studies addressed on- and off-task behavior (see Table 1). Of these six, three used SSEDs (Darch & Thorpe, 1977; Fishbein & Wasik, 1981; Lannie & McCurdy, 2007) and three (Dion et al., 2011; Leflot et al., 2010; Leflot et al., 2013)

Table 1

Study Outcomes by Dependent Variable

Dependent variable	Study author(s)	Design	Effect size (*d*)	PND%	PAND%
Disruptive Behavior	Darveaux (1984)	SSED		Child 1 = 100 Child 2 = 100	
	Donaldson, Vollmer, Krous, Downs, and Berad (2011)	SSED		Experimenter: Class 1 = 100 Class 2 = 100 Class 3 = 100 Class 4 = 91.66 Class 5 = 100 Teacher: Class 1 = 100 Class 2 = 91.6 Class 3 = 88.88 Class 4 = 78.97 Class 5 = 100	Experimenter: Class 1 = 100 Class 2 = 100 Class 3 = 100 Class 4 = 96.15 Class 5 = 100 Teacher: Class 1 = N/A Class 2 = 96 Class 3 = 93.75 Class 4 = 90.9 Class 5 =100
	Fishbein and Wasik (1981)	SSED		Library = 100 Classroom = 0	
	Lannie and McCurdy (2007)	SSED		27.27	
	McCurdy, Lannie, and Barnabas (2009)	SSED		Lunch periods: K, 3rd = 100 1st, 2nd = 100 4th, 6th = 33.33	Lunch Periods: K, 3rd = 100 1st, 2nd = 100 4th, 6th = 95
	Medland and Stachnik (1972)	SSED		90	89.72
	Ruiz-Olivares, Pino, and Herruzo (2010)	SSED		Teacher 1: 100 Teacher 2: 100	
	Salend, Reynolds, and Coyle (1989)	SSED		100	100

(continued)

Table 1 (continued)

Dependent variable	Study author(s)	Design	Effect size (*d*)	PND%	PAND%
Off-task/On-task	Darch and Thorpe (1977)	SSED		100	
	Dion et al. (2011)	Exp	Attentive = .81 Inattentive = 1.22		
	Fishbein and Wasik (1981)	SSED		Library = 85.71 Classroom 28.57	
	Lannie and McCurdy (2007)	SSED		90.91	
	Leflot, van Lier, Onghena, and Colpin (2010)	Exp	Wave 2 = 0.61 Wave 4 = 0.22		
	Leflot, van Lier, Onghena, and Colpin (2013)	Exp	Wave 4 = −0.47 (for low on-task)		
Aggression	Kellam, Ling, Merisca, Brown, and Ialongo (1998)	Exp	Overall = 0.11; (up to *d* = 0.32 at higher levels of aggression)		
	Kellam, Rebok, Ialongo, and Mayer (1994)	Exp	−0.39		
	Kleinman and Saigh (2011)	SSED		77.78	
	Leflot et al. (2013)	Exp	0.48 (low on-task)		
	Saigh and Umar (1983)	SSED		91.67	95.83
Talking Out	Barrish, Saunders, and Wolf (1969)	SSED		Math = 100 Reading = 100	Math = 100 Reading = 100
	Leflot et al. (2010)	Exp	Wave 2 = −.62 Wave 4 = −.29		
	Saigh and Umar (1983)	SSED		75	100
	Salend et al.(1989)	SSED		Class A = 100 Class B = 87.25	Class A = 100 Class B = 89.29
Out-of-seat	Barrish et al. (1969)	SSED		Math = 100 Reading = 100	Math = 100 Reading = 100
	Kleinman and Saigh (2011)	SSED		100	
	Leflot et al. (2010)	Exp	Wave 2 = .13 Wave 4 = .08		
	Saigh and Umar (1983)	SSED		66.67	83.33

(continued)

Table 1 (continued)

Dependent variable	Study author(s)	Design	Effect size (*d*)	PND%	PAND%
Peer Acceptance/ Rejection (acc/rej)	Leflot et al. (2013)	Exp	0.41 (rej) Low on-task		
	Witvliet, van Lier, Cuijpers (2009)	Exp	0.34 (acc)		
Rule Violations	Tanol, Johnson, and McComas (2010)	SSED		Class 1 = 100 Class 2 = 100	
	Bostow and Geiger (1976)			100	
Antisocial Negative Behaviors	McGoey, Schneider, Rezzetano, Prodan, and Tankersley (2010)	SSED		Class 1 = 56.25 Class 2 = 31.58 Class 3 = 68.42	
Appropriate and Inappropriate Social Interactions	Patrick, Ward, and Crouch (1998)	SSED		Inappropriate = 100 Appropriate = 100	Inappropriate = 100 Appropriate = 100
Externalizing Behavior	Witvliet et al. (2009)	Exp	0.45		
Swearing or Negative Comments	Salend et al. (1989)	SSED		100	100

Note. PND% = percent nonoverlapping data points; PAND% = percent all nonoverlapping data points; Exp = experimental design; SSED = single subject experimental design.

used experimental designs to study the effects of the GBG. The researchers for four studies specifically measured on-task behavior (Darch & Thorpe, 1977; Dion et al., 2011; Leflot et al., 2010; Leflot et al., 2013) and two measured off-task behavior (Fishbein & Wasik, 1981; Lannie & McCurdy, 2007). Only one study (Fishbein & Wasik, 1981) found a limited effect of the GBG intervention on off-task behavior. For the other studies, the GBG realized moderate to large effects.

Aggression

Five studies addressed aggressive behavior. The definition for aggression was physical contact such as hitting, kicking, tapping, tripping, pinching, throwing objects in the classroom, and destroying the property of others. Two studies (Kleinman & Saigh, 2011; Saigh & Umar, 1983) used SSEDs and three studies (Kellam et al., 1994; Kellam et al., 1998; Leflot et al., 2013) had experimental designs. The researchers who conducted the two SSED studies suggested that the GBG was a moderately to highly effective intervention for reducing aggression. All three experimental studies demonstrated that use of the GBG had modest effects on aggression. The study by Kellam et al. (1994) also indicated that the GBG had a greater reductive effect in more aggressive children, whereas Leflot et al.'s (2013) findings suggested that the modest effect was only for students who had low baseline on-task behavior and no effect on students who had high on-task behavior at baseline.

Talking Out

Four studies addressed talking out which defined primarily as talking without permission. Three of these four studies used SSEDs to study the effects of GBG implementation on talking out behavior (Barrish et al., 1969; Saigh & Umar, 1983; Salend et al., 1989). Overall, GBG implementation had moderately positive effects on the reduction of talking out. Barrish et al. (1969) suggested that the GBG was highly effective for reducing talking out behavior. Salend et al. (1989) indicated that the GBG was highly effective for one class and moderately effective for another class. Saigh and Umar (1983) and Leflot et al. (2010) found that the GBG was moderately effective for reducing talking out.

Out of Seat

Four studies focused on out-of-seat behavior. Out-of-seat behavior was defined as leaving one's seat or seated position without permission. Three studies employed SSEDs (Barrish et al., 1969; Kleinman & Saigh, 2011; Saigh & Umar, 1983) and one study (Leflot et al., 2010) used an experimental design. Barrish et al. (1969) and Kleinman and Saigh (2011) indicated that the GBG was a very effective intervention to reduce out-of-seat behavior. Saigh and Umar (1983) found the GBG to be a mildly effective intervention based on PND and a moderately effective intervention for out-of-seat behavior according to PAND. Leflot et al. (2010) indicated a near null effect of the GBG intervention on out-of-seat behavior.

Peer Acceptance and Rejection

Two studies (Leflot et al., 2013; Witvliet, van Lier, & Cuijpers, 2009) addressed peer acceptance and/or rejection. The definition for peer acceptance was that

other students liked the student. Peer rejection referred to being liked least by classmates. Leflot et al. (2013) asked students to nominate all other students that they liked least (rejection). Witvliet et al. (2009) requested the opposite and asked students to nominate all other students that they liked most (acceptance). Both studies used experimental designs. Findings suggested that the GBG had a modest effect on increasing acceptance and decreasing rejection. In the case of decreasing peer rejection, Leflot et al.'s (2013) findings suggested that this was particularly true for students with low levels of on task behavior at baseline.

Rule Violations

Two studies (Bostow & Geiger, 1976; Tanol et al., 2010), both SSEDs, concerned rule violations in the classroom. Rule violations were generally defined as not following rules or engaging in behaviors against classroom expectations. Although the specific behaviors of interest for other studies could also be considered rule violations, those other behaviors were more specific than the definitions used in these two studies. These researchers found the GBG to be a highly effective intervention for reducing rule violations.

Antisocial/Negative Behavior

One study (McGoey, Schneider, Rezzetano, Prodan, & Tankersley, 2010) focused on antisocial/negative behavior using a SSED. Antisocial/negative behavior was defined as a composite of several behaviors including negative social interactions, off-task behavior, and tantrumming. Their findings suggested that the GBG is an ineffective intervention against these antisocial/negative behaviors.

Appropriate and Inappropriate Social Interactions

One study (Patrick, Ward, & Crouch, 1998) used an SSED to address appropriate and inappropriate social interactions. These behaviors were incompatible categories of behavior where appropriate behavior was defined as supportive verbal, physical, or gestural acts. Inappropriate social interactions were defined as aggressive verbal, physical, or gestural acts. In this study, the GBG appeared to be a highly effective intervention for increasing appropriate interactions and decreasing inappropriate interactions.

Externalizing Behavior

Using an experimental design, one study (Witvliet et al., 2009) addressed the effect of the GBG on externalizing behavior. Externalizing behavior referred to oppositional and conduct problems. Using teacher rating scales researchers measured students' oppositional and conduct problems as observed by teachers, findings revealed that use of the GBG had a modest effect on externalizing behavior.

Swearing/Negative Comments to Others

Using an SSED, Salend et al. (1989) used the GBG in an attempt to reduce swearing and negative comments among 19 high school students with emotional disturbance. Swearing referred to verbal statements or gestures pertaining to body parts designed for sexual activity or waste elimination, uncomplimentary references to others' parentage. Negative comments referred to negative verbal

Table 2

Parameter Estimates (With Standard Errors) From HLM Meta-analysis of GBG SSEDs

	Intercept	Baseline slope	Immediate treatment effect	Treatment effect on trend
Fixed effects	51.88** (8.05)	−0.51 (0.40)	−20.38* (7.30)	−0.03 (0.48)

Note. SSED = single subject experimental design; HLM = hierarchical linear modeling; GBG = Good Behavior Game.
*$p < .01$. **$p < .001$.

comments, complaints about assignments, or complaints about the instruction. Salend et al. (1989) suggested that the GBG was a highly effective intervention for reduction of swearing and negative comments.

Research Question 1: Effects of the GBG

The average Cohen's *d*, calculated from all group design studies, revealed a moderate effect ($d = .50$) of the GBG intervention on challenging behaviors in classroom and school settings. Using the data extracted from the graphs from the SSEDs, the HLM analysis revealed that the overall baseline mean (intercept) for challenging behavior was 51.88%. Across studies the immediate treatment effect (β_{2jk}) was −20.38%. These results indicated that a high rate of challenging behavior during the baseline phase was evident and an immediate decrease in the behavior occurred with introduction of the treatment. Both the baseline mean ($p < .001$) and immediate treatment effect ($p < .01$) were statistically significant. The treatment effect on trend (β_{3jk}) also decreased slightly, by .03% during the treatment phase; however, this effect was not statistically significant, $p > .05$ (see Table 2).

Research Question 2: Outcome Differences by Study Features

The GBG was found to be a highly effective intervention across a range of challenging classroom and school behaviors (see Table 1). According to Cohen's *d*, PND, and PAND, 10 studies indicated a null, modest, or mild effect for at least one case (Fishbein & Wasik, 1981; Kellam et al., 1994; Kellam et al., 1998; Lannie & McCurdy, 2007; Leflot et al., 2010; Leflot et al., 2013; McCurdy et al., 2009; McGoey et al., 2010; Saigh & Umar, 1983; Witvliet et al., 2009). Through this research question we examined some potential reasons for these differing results, including a review of characteristics including fidelity, who served as the intervention agent, the type of training provided to the GBG interventionist, setting, intervention duration, and rewards provided to students. Table 3 summarizes these findings.

Fidelity

Fidelity was reported in eight studies (Dion et al., 2011; Donaldson, Vollmer, Krous, Downs, & Berad, 2011; Lannie & McCurdy, 2007; Leflot et al., 2010; Leflot et al., 2013; McCurdy et al., 2009; Salend et al., 1989; Tanol et al., 2010). Fourteen studies did not report fidelity results. Fidelity scores were near or above 80% in all

Table 3

Features of GBG Studies

Study	Teacher implement	Interventionist training				Setting (elem/sec)	Duration (in days)	Rew (Y/N)	Fid	Eff
		Lecture	Lecture and feedback	Model	NR					
Barrish et al. (1969)	●				●	Elem	56	Y	N	High, High
Bostow and Geiger (1976)	●	●				Elem	25	Y	N	High
Darch and Thorpe (1977)	Student teacher				●	Elem	27	Y	N	High
Darveaux (1984)	●		●			Elem	20	Y	N	High
Dion et al. (2011)	●		●			Elem	120	Y	Y	High
Donaldson et al. (2011)	●			●		Elem	Unk	Y	Y	High
Fishbein and Wasik (1981)	Librarian and teacher				●	Elem	65	Y (librarian only)	N	High, Null
Kellam et al. (1998)	●	●				Elem	Unk	N	N	Null, Modest
Kellam et al. (1994)	●	●				Elem	Unk	N	N	Modest
Kleinman and Saigh (2011)	●	●				Sec	21	Y	N	High-Mod
Lannie and McCurdy (2007)	●		●			Elem	17	Y	Y	Ineff-High
Leflot et al. (2010)	●		●			Elem	320	N	Y	Null-Mod
Leflot et al. (2013)	●		●			Elem	320	N	Y	Modest-Modest
McCurdy et al. (2009)	Lunch assistants	●				Elem	20	Y	Y	Ineff, High
McGoey et al. (2010)	●				●	Elem	37	Y	N	Ineff-Mild
Medland and Stachnik (1972)	●				●	Elem	55	Y	N	Moderate
Patrick et al. (1998)	●				●	Elem	20	N	N	High
Ruiz-Olivares et al. (2010)	●	●			●	Elem	30	Y	N	High
Saigh and Umar (1983)					●	Elem	25	N	N	Mild-High
Salend et al. (1989)	●				●	Sec	14	Y	Y	Moderate-High
Tanol et al. (2010)	●		●			Elem	40	Y	Y	High
Witvliet et al. (2009)	●		●			Elem	Unk	Y	N	Modest

Note. Teacher Implement refers to whether a teacher was the primary implementer of the GBG intervention. NR under interventionist training indicates: not recorded or not given. In the setting column, elem refers to elementary school and sec refers to secondary school (middle school/high school). Rew refers to whether rewards were indicated and/or described. Fid refers to whether fidelity was measured. Eff refers to the effectiveness. A dash (-) between two effectiveness levels (e.g., mod-high) describes effectiveness across two DVs. A comma (,) between two effectiveness levels (mod, high) indicates effectiveness across cases on one DV.

except for one (Donaldson et al., 2011) of the eight studies that reported fidelity outcomes. Donaldson et al. (2011) measured fidelity and determined that fidelity averaged 60%; however, the lower fidelity in this study did not appear to affect the intervention outcomes as PND scores were found to be more than 90%. Two groups of researchers (McGoey et al., 2010; Ruiz-Olivares et al., 2010), both conducting SSEDs, indicated that they did not formally assess fidelity, but both acknowledged the lack of fidelity assessment as a limitation. One additional study (Patrick et al., 1998) suggested training to ensure fidelity but did not measure fidelity.

Intervention Agent

School staff served as intervention agents in the majority of studies (n = 21) and teachers were the most common (n = 19). In one study (Darch & Thorpe, 1977), a student teacher was the intervention agent and the GBG was found to be highly effective for off-task and out-of-seat behavior. In another study (Fishbein & Wasik, 1981), the school librarian was the primary implementer of the intervention rather than the classroom teacher even though effects were measured in both the classroom and the library. The GBG appeared to reduce disruptive and off-task behavior in the library, but not in the classroom. Finally, McCurdy et al. (2009) focused on behavior in the school cafeteria where lunchtime supervision staff implemented the GBG intervention and, overall, found large effects.

Thirteen articles mentioned interventionist GBG training or training materials. Seven studies (Darveaux, 1984; Dion et al., 2011; Lannie & McCurdy, 2007; Leflot et al., 2010; Leflot et al., 2013; Tanol et al., 2010; Witvliet et al., 2009) referred to training as including a combination of a lecture and or follow-up consultation and feedback with teachers upon implementation. Donaldson et al. (2011) also indicated that training procedures included the experimenter implementing the GBG in the presence of the teacher prior to the teacher assuming responsibility for the GBG. See Table 3 for a summary of the training methods used in these studies.

Of the 10 articles that reported limited to null effects on various outcome variables, two did not indicate how intervention agents were trained (Fishbein & Wasik, 1981; McGoey et al., 2010). McCurdy et al. (2009) did indicate training for the interventionist. However, the training appeared to be quite brief as interventionists participated in one 90-minute training procedure that included role-play and feedback. In one study (Leflot et al., 2010), GBG effects were minimal on one of the DVs (out-of-seat behavior); however, interventionist training appeared to be somewhat intensive as intervention agents were provided manuals, three half-day trainings, and 10 one-hour observations of their implementation.

Duration

Duration data were available, calculated, or estimated for 18 of 22 of the studies. Seven articles with duration information (Fishbein & Wasik, 1981; Lannie & McCurdy et al., 2007; Leflot et al., 2010; Leflot et al., 2013; McCurdy et al., 2009; McGoey et al., 2010; Saigh & Umar, 1983) had modest, questionable, or null effects on at least one DV. The duration data for the highly or moderately effective studies were compared with the duration data from studies with modest or questionable effects. For the studies with high or moderate effects, the median

number of intervention days was 26 (range = 14-120). The median for duration for studies with modest, questionable, or null effects was 37 days (range = 17-320). This suggests that longer duration of GBG implementation does not necessarily mean better outcomes.

Setting

All 22 articles described the instructional context in which the GBG study took place. Twenty of the studies were conducted in elementary schools and classrooms. Two studies were conducted in secondary school settings—that is, a ninth-grade history class (Kleinman & Saigh, 2011) and a residential setting with high school students with emotional disturbance (Salend et al., 1989). Overall, school setting did not appear to affect GBG findings as moderate to large effects were found across elementary and secondary settings.

Rewards

The GBG naturally provides teachers with a vehicle through which to administer rewards to their students. Rewards in the GBG serve to increase students' use of appropriate behaviors versus more challenging behaviors. Rewards or winning were referred to in 16 articles (see Table 3). Most articles indicated if verbal, tangible, or social/activity rewards were used. However, Saigh and Umar (1983) did not mention the type of rewards given beyond saying they used rewards. Three articles indicated verbal praise (Darch & Thorpe, 1977; Dion et al., 2011; McCurdy et al., 2007). Tangibles were the most commonly used type of reward (n = 14; Barrish et al., 1969; Bostow & Geiger, 1976; Darveaux, 1984; Donaldson et al., 2011; Fishbein et al., 1981; Kellam et al., 1994; Kleinman & Saigh, 2011; Lannie & McCurdy, 2007; McGoey et al., 2010; Medland & Stachnik, 1972; Ruiz-Olivares et al., 2010; Salend et al., 1989; Tanol et al., 2010; Witvliet et al., 2009), and were often combined with verbal praise.

Studies that utilized tangible rewards appear to have had the highest effects as 9 of the 14 studies that used tangibles had high or moderate effects. Five articles (Fishbein & Wasik, 1981; Kellam et al., 1994; Lannie & McCurdy, 2007; McGoey et al., 2010; Witvliet et al., 2010) indicated use of tangible rewards but demonstrated no effect or modest or questionable effects. In these five studies, rewards were delivered daily (Fishbein & Wasik, 1981; McGoey et al., 2010; Witvliet et al., 2009), daily and weekly (Kellam et al., 1994), or the schedule was unspecified (Lannie & McCurdy, 2007). Two of these studies (Lannie & McCurdy, 2007; McGoey et al. 2010) made use of edibles for rewards, particularly candy. McGoey et al. (2010) appeared to offer a variety of options but the teacher initially selected these rewards.

In total, only four studies made use of preference assessments (Kleinman & Saigh, 2011; Lannie & McCurdy, 2007; Saigh & Umar, 1983; Salend et al., 1989). Three of the studies that indicated use of preference assessments had moderate or large effects (Kleinman & Saigh, 2011; Saigh & Umar, 1983; Salend et al., 1989) and one suggested modest effects (Lannie & McCurdy, 2007). Perhaps one of the most poignant findings with regard to the GBG and reward use is that one study (Fishbein & Wasik, 1981) found that the GBG had a large effect when a reward was used and a null effect when not used. Rewards used with the GBG appear be

a critical component with regard to increasing students' appropriate behaviors and simultaneously decreasing challenging behavior.

Discussion

By evaluating Cohen's *d*, PND, PAND, and results of an HLM analysis, we concluded that, overall, the GBG appears to have a moderate to large effect on challenging behaviors in school and classroom setting. Results of the HLM analysis for SSEDs suggested that at baseline challenging behavior was high and GBG implementation resulted in an immediate and significant change in level. The GBG also appeared to have a continued effect throughout the intervention phase as the trend continued to evidence an extremely slight decrease in challenging behavior. Use of HLM to evaluate the overall effect of SSEDs proved useful as all data points from SSEDs could be used and simultaneously analyzed, neither of which is possible using overlapping data metrics such as PND or PAND.

Effectiveness of the GBG

An examination of effects at the individual study level revealed PNDs and PANDs with moderate to large effects overall. Of 45 PNDs across 16 studies, 37 were in the moderate to large effect range and 31 PNDs were over 90%. Eight PNDs from five studies (Fishbein & Wasik, 1981; Lannie & McCurdy, 2007; McCurdy et al., 2009; McGoey et al., 2010; Saigh & Umar, 1983) were indicative of mild or null effects; however, two of these (McCurdy et al., 2009; Saigh & Umar, 1983) were artifacts of the PND metric, as PAND suggested moderately to highly effective interventions. In these cases, PND's overreliance on outlying baseline data points appeared to be a problem. These results indicate that only 6 of 45 PNDs reflected mild or null effects. Interestingly, five of these six PNDs were for DVs that consisted of a combination of multiple behaviors, for example, disruptive behavior and antisocial/negative behaviors. This observation may speak to the need for researchers to improve their operationalization of the DVs and measure effectiveness with greater specificity regarding behaviors and classes of behavior.

Across the group design studies, Cohen's *d* tended to be in the moderate range. Shadish, Rindskopf, and Hedges (2008) suggested that these effect size differences (i.e., between Cohen's *d* and PND/PAND) are somewhat common and unsurprising because SSEDs contain within-subject variability as the primary source of variation. On the other hand, experimental designs produce between subject variability. Data from an individual are expected to be much less variable than data based on a group. Another possible explanation might be simply that the single subject studies for which PND and PAND were calculated were typically based on observational data, whereas large *N*, experimental studies used rating scales more frequently. Although the use of rating scales and surveys allows researchers to efficiently obtain a large quantity of information, these data may be vulnerable to various sources of error, such as the error of recency, where a rater may remember only the most recent events or behaviors exhibited by a student.

Although large values for Cohen's *d* were not commonly found, Dion et al. (2011) did find large effects of the GBG on students' on-task behavior. In this study a larger effect was found for the GBG's effect on inattentive students'

attention compared with the attention of attentive students, although *d* for both groups was large. This finding suggests that the GBG helped both groups of students to be more attentive but that the group that had more room to improve made more progress. Leflot et al. (2013) also found that the GBG had greater effects (on on-task behavior, aggression, and peer rejection) for students who had low levels of on-task behavior at baseline. Kellam et al. (1998) found an overall null effect in terms of effect size, but when they analyzed data by groups of students based on baseline aggression levels, they found that the GBG had modest effects on the aggression levels of students with higher initial levels of aggression. These findings have implications for use of the GBG among students with aggression and attention problems and various disabilities—that is, the GBG may be of particular use among students with attention- and aggression-related school problems.

Another interesting finding from analysis of the group design studies comes from the work by Leflot et al. (2010). Leflot et al. found a moderate effect of the GBG on talking out and on-task behavior; however, stronger effects were observed at the first post intervention period (Wave 2) than at the end of the following year with the same participants (Wave 4). These differences between waves are probably best explained by considering that there may be floor effects for DVs. For example, if students talk out at 0% of the time by Wave 2, they cannot improve any further which will eliminate subsequent effects. Findings from Leflot et al. (2010) are also reflective of the HLM findings from analysis of SSEDs in our study, as we found that the GBG effects were immediate and largely stable with only a very slight decrease over time. Overall, the GBG appears to be effective at reducing a variety of challenging behaviors in school settings. Disruptive behaviors including off-task, aggression, talking out, and out of seat, behaviors could potentially impede teaching and learning. A simple management procedure, such as the GBG, may facilitate increased time devoted to teaching and learning.

Impact of Study Features on Outcomes

We examined various features of every study to assess the GBG's impact on the outcome. One of these variables was fidelity. When considering intervention studies, fidelity is an important component as low fidelity scores threaten internal validity. After comparing findings for articles that contained fidelity information with those that did not contain that information, we found that the lack of fidelity data did not appear to have an effect on outcomes, as modest and large findings were found across articles with the information and across articles without the information. Nonetheless, the lack of fidelity information was a surprising finding, particularly given that measuring fidelity is a common research standard.

We compared studies with high/moderate effects as given by Cohen's *d* or PND with studies with modest, questionable or null outcomes to understand how various study features (interventionist experience and training, setting, duration, and reward use) contributed to magnitude of effect. Given that most studies provided some level of training for interventionists, it is difficult to determine how training affected outcomes, as there did not appear to be any particular differences in findings that could be associated with type of training provided. Researchers implemented the GBG primarily in elementary general education settings with moderate to high effects. Those GBG implementations conducted in secondary

settings were also highly effective. Findings for duration suggested that GBG implementation did not require lengthy intervention, as studies with modest or null effects had longer durations than studies with moderate or large effects. This finding is also consistent with results of our HLM analysis, which indicated that GBG results in an immediate drop in challenging behavior, but only slight additional change over time.

Of the study features studied here, use of rewards appears to have some effect on magnitude of outcome. It appears that where modest or null effects were found, rewards were not used or were used in a limited way. For example, Fishbein and Wasik (1981) found that the GBG was highly effective in the library where the librarian administered rewards to students. The GBG was not effective in the classroom as the librarian was not present and rewards were not administered. Another issue concerns making the rewards of interest to the students. For example, when the teacher chose the rewards (McGoey et al., 2010), the students may not have a clear understanding of what rewards they would earn. Also, as highlighted, many of the researchers did not use preference assessments to identify rewards preferred by the students. Of the four articles that reported use of preference assessments, three indicated large or moderate effects. These strong effects support the use of preference assessments with GBG implementation.

Limitations

Like all studies this review has some limitations to address. First, this study relied on the already published peer reviewed literature. This decision means that sound implementations of the GBG conducted for dissertation or thesis research may have been missed. Furthermore, we were only able to use the information provided in the included articles. For example, information on study duration was not included in all articles. It is possible that these omissions introduced some error into our analysis of the effect of duration on GBG outcomes. Second, several studies were eliminated from this review because they did not include challenging behaviors in school and classroom settings as dependent variables. These additional studies illustrate merits of the GBG, for example, the relationship between the GBG and later substance abuse or depression. Also, studies were only included when the effect of the GBG could be isolated. Finally, a number of studies were excluded as they were not written in English (e.g., Pérez, Rodríguez, De la Barra, & Fernández, 2005, which was written in Spanish) and we were not able to code them. It is likely that these studies would also be useful in evaluating the GBG effects on challenging behaviors in school and classroom settings.

Implications for Practice

Several implications for practice must be highlighted as a result of this review. First, the GBG appears to be an effective intervention to address a variety of challenging behaviors that could potentially cause an interruption to the teaching and learning process in the classroom. With GBG implementation, teachers may be able to spend more time on teaching and less time responding to behavioral incidents in the classroom. This outcome is important given the academic growth requirements that schools face (Vannest et al., 2009). Additionally, the GBG appears to have immediate and moderate to large effects over short periods of

time. The fact that the GBG shows effects after a short time may be particularly important for teachers in classrooms where challenging behavior is frequent and teachers are struggling with classroom management. In these situations, GBG use may facilitate changes in student behavior so that teaching and learning can take place.

Another important implication concerns the types of behaviors addressed with the GBG. Through our review, we found that most GBG research focused on externalizing, challenging behaviors. This finding is not surprising as this type of behavior is generally characterized as noncompliant, disruptive, and often aggressive. Students with this type of behavior may be more noticeable or present greater concern for teachers compared with other behaviors or needs (Cullinan, Evans, Epstein, & Ryser, 2003). Externalizing behavior problems are commonly referred to the school office for intervention (McIntosh, Campbell, Carter, & Zumbo, 2009), as they are likely to interfere with teaching, learning, safety, and/or relationships. Harrison et al. (2012) found that teachers consistently cite student distractibility as a major concern. With the GBG demonstrated as effective for decreasing off-task behaviors and increasing the amount of time that students pay attention, teachers who are concerned with student distractibility may also view its use positively and aim to incorporate it into their classroom management systems. The effectiveness of the GBG on externalizing and off-task behaviors offers promise for teaching and learning.

As this review demonstrates, the GBG has been implemented by individuals in a variety of school roles such as classroom teachers, student teachers, librarians, and lunchtime staff. The different school roles of the interventionists highlight the ease in which the GBG can be implemented under a variety of conditions. Additionally, the relatively brief training for interventionists evidenced in these articles suggests that the GBG can be used successfully without extensive GBG training. When teachers consider using the GBG as part of their classroom management procedures they should consider how they will reward desired behavior and whether they will conduct preference assessments to understand student interests. Given the results synthesized here, the GBG might be considered as a promising practice for the classroom.

Implications for Future Research

This review has illustrated the magnitude of the effect of the GBG on various school and classroom behavioral outcomes. We examined the effects of various GBG study features on the associated outcomes. One of the quality indicators for SSEDs is the inclusion of social validity measurement. Future researchers might examine the effects of high or low social validity in SSEDs on GBG outcomes as well. Even though researchers have studied the GBG numerous times, there are still gaps in the literature in terms of implementation and effect on problem behavior across grade levels, school settings, and disability status. Since most studies reviewed here focused on elementary students, further use in secondary schools to determine effects on problem behaviors such as attendance for older adolescents might be useful. Additionally, use of the GBG to address internalizing behaviors is another avenue for exploration. Future studies should also provide detailed information on study participants and various training procedures, as this

information was often lacking in studies that met inclusion criteria. Finally, with the academic achievement requirements required by NCLB, future research might consider the effects of the GBG on academic outcomes.

References

References marked with an asterisk indicate studies included in the review.

Alresheed, F., Hott, B., & Bano, C. (2013). Single subject research: A synthesis of analytic methods. *Journal of Special Education Apprenticeship*, *2*, 1–18.

*Barrish, H., Saunders, M., & Wolf, 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.

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: Promotion of self-regulation and the prevention of early school failure. *Development and Psychopathology*, *20*, 899–911. doi:10.1017/S0954579408000436

Bohannon, H., & Wu, M.-J. (2011). Can prevention programs work together? An example of school-based mental health with prevention initiatives. *Advances in School Mental Health Promotion*, *4*(4), 35–46.

*Bostow, D., & Geiger, G. (1976). Good behavior game: A replication and systematic analysis with a second grade class. SALT: *School Applications of Learning Theory*, *8*(2), 18–27.

Boyle, M. A., Samaha, A. L., Rodewald, A. M., & Hoffman, A. N. (2012). Evaluation of the reliability and validity of GraphClick as a data extraction program. *Computers in Human Behavior*, *29*, 1023–1027. doi:10.1016/j.chb.2012.07.031

Cameron, C., Connor, C., Morrison, F., & Jewkes, A. (2008). Effects of classroom organization on letter-word reading in first grade. *Journal of School Psychology*, *6*, 173–192. doi:10.1016/j.jsp.2007.03.002

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Cullinan, D., Evans, C., Epstein, M., & Ryser, G. (2003). Characteristics of emotional disturbance of elementary school students. *Behavioral Disorders*, *28*, 94–110.

*Darch, C., & Thorpe, H. (1977). The principal game: A group consequence procedure to increase classroom on-task behavior. *Psychology in the Schools*, *14*, 341–347. doi:10.1002/1520-6807(197707)14:3<341::AID-PITS2310140315>3.0.CO;2-3

*Darveaux, D. (1984). The Good Behavior Game plus merit: Controlling disruptive behavior and improving student motivation. *School Psychology Review*, *13*, 510–514.

*Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., & Dupéré, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science*, *12*, 70–79. doi:10.1007/s11121-010-0182-5

*Donaldson, J. M., Vollmer, T. R., Krous, T., Downs, S., & Berad, K. P. (2011). An evaluation of the Good Behavior Game in kindergarten classrooms. *Journal of Applied Behavior Analysis*, *44*, 605-609. doi:10.1901/jaba.2011.44-605

Elswick, S., & Casey, L. B. (2012). The Good Behavior Game is no longer just an effective intervention for students: An examination of the reciprocal effects on teacher behaviors. *Beyond Behavior*, *21*, 36–46.

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., & Wasik, B. (1981). Effect of good behavior game on disruptive library behavior. *Journal of Applied Behavior Analysis*, *14*, 89–93. doi:10.1901/jaba.1981.14-89

Harrison, J., Vannest, K., Davis, J., & Reynolds, C. (2012). Common problem behaviors of children and adolescents in general education classrooms in the United States. *Journal of Emotional and Behavioral Disorders*, *20*, 55–64. doi:10.1177/1063426611421157

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446. (2004). Retrieved from http://idea.ed.gov/explore/view/p/%2Croot%2Cstatute%2C

Kaufman, J. S., Jaser, S. S., Vaughan, E. L., Reynolds, J. S., Di Donato, J., Bernard, S. N., & Hernandez-Brereton, M. (2010). Patterns in office discipline referral data by grade, race/ethnicity, and gender. *Journal of Positive Behavior Interventions*, *12*, 44–54. doi:10.1177/1098300708329710

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

*Kellam, S. G., Ling, X., Merisca, R., Brown, C. H., & Ialongo, N. (1998). The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Development and Psychopathology*, *10*, 165–185. doi:10.1017/S0954579498001564

*Kellam, S., Rebok, G., Ialongo, N., & Mayer, L. (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. doi:10.1111/j.1469-7610.1994.tb01161.x

Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, *102*, 741–756. doi:10.1037/a0019237

*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. doi:10.1177/0145445510392213

Kratochwill, T., Stoiber, K., Christenson, S., Durlak, J., Levin, J., Talley, R., & Shadish, W. R. (2002). In T. R. Kratochwill & K. C Stoiber (Eds.), *Procedural and coding manual for review of evidence-based interventions* (Task force on the Evidence Based Interventions in School Psychology Sponsored by Division 16 of the American Psychological Association and Society for the Study of Psychology). Retrieved from http://www.indiana.edu/~ebi/documents/_workingfiles/EBImanual1.pdf

*Lannie, A. L., & McCurdy, B. L. (2007). Preventing disruptive behavior in the urban classroom: Effects of the Good Behavior Game on student and teacher behavior. *Education & Treatment of Children*, *30*, 85–98. doi:10.1353/etc.2007.0002

*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.

*Leflot, G., van Lier, P. A. C., Onghena, P., & Colpin, H. (2013). The role of children's on-task behavior in the prevention of aggressive behavior development and peer rejection: A randomized controlled study of the Good Behavior Game in Belgian

elementary classrooms. *Journal of School Psychology*, *51*, 187–199. doi:10.1016/j.jsp.2012.12.006

Little, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS® System for mixed models*. Cary, NC: SAS Institute.

Mastropieri, M. A., Scruggs, T. E., Bakken, J. P., & Whedon, C. (1996). Reading comprehension: A synthesis of research in learning disabilities. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Advances in learning and behavioral disabilities* (*Vol. 10*, Part B, pp. 201–227). Greenwich, CT: JAI Press.

*McCurdy, B., 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*, 39–54. doi:10.1016/j.jsp.2008.09.003

*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. doi:10.1080/15377903.2010.495916

McIntosh, K., Campbell, A., Carter, D., & Zumbo, D. (2009). Concurrent validity of office discipline referrals and cut points used on schoolwide positive behavior support. *Behavioral Disorders*, *34*, 100–113.

*Medland, M., & Stachnik, T. (1972). Good behavior game: A replication and systematic analysis. Journal of Applied Behavior Analysis, *5*, 45–51.

Nagler, E. M., Rindskopf, D. M., & Shadish, W. R. (2008). *Analyzing data from small N designs using multilevel models: A procedural handbook*. Washington, DC: U.S. Department of Education.

Nelson, J. R., Maculan, A., Roberts, M. L., & Ohlund, B. J. (2001). Source of occupational stress for teachers of students with emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders*, *9*, 123–131. doi:10.1177/106342660100900207

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percent of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education*, *40*, 194–204.

*Patrick, C., Ward, P., & Crouch, D. (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.

Pérez, V., Rodríguez, J., De la Barra, F., & Fernández, A. M. (2005). Efectividad de una estrategia conductual para el manejo de la agresividad en escolares de enseñanza básica [Effectiveness of a behavioral strategy of aggression management in elementary school children]. *Psykhe*, *14*, 55–62. doi:10.4067/S0718-22282005000200005

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear modeling: Applications and data analysis procedures* (2nd ed.). Thousand Oaks, CA: Sage.

*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. doi:10.1002/pits.20523

*Saigh, P., & Umar, A. (1983). The effects of the good behavior game on the disruptive behavior of Sudanese elementary school students. *Journal of Applied Behavior Analysis*, *16*, 339–344. doi:10.1901/jaba.1983.16-339

*Salend, S., Reynolds, C., & Coyle, E. (1989). Individualizing the good behavior game across type and frequency of behavior with emotionally disturbed adolescents. *Behavior Modification*, *13*, 108–126. doi:10.1177/01454455890131007

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*, 24–33.

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*, 188–196.

Spaulding, S. A., Irvin, L. K., Horner, R. H., May, S. L., Emeldi, M., Tobin, T. J., & Sugai, G. (2010). Schoolwide social-behavioral climate, student problem behavior, and related administrative decisions: Empirical patterns from 1,510 schools nationwide. *Journal of Positive Behavior Interventions*, *12*, 69–85. doi:10.1177/1098300709332345

*Tanol, G., Johnson, L., & McComas, J. (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. doi:10.1016/j.jsp.2010.06.001

Tingstrom, D., Sterling-Turner, H., & Wilczynski, S. (2006). The Good Behavior Game: 1969-2002. *Behavior Modification*, *30*, 225–253. doi:10.1177/0145445503261165

U.S. Department of Education. (2000). *Annual report to Congress on the implementation of the Individuals with Disabilities Act*. Washington, DC: Office of Special Education and Rehabilitative Services.

Van den Noortgate, W., & Onghena, P. (2003). Combining single case experimental data using hierarchical linear modeling. *School Psychology Quarterly*, *18*, 325-346. doi:10.1521/scpq.18.3.325.22577

Vannest, K., Temple-Harvey, K., & Mason, B. (2009). Adequate yearly progress for students with emotional and behavioral disorders through research-based practices. *Preventing School Failure*, *53*, 73–84.

Walter, H. J., Gouze, K., & Lim, K. G. (2006). Teachers' beliefs about mental health needs in inner city elementary schools. *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*, 61–68. doi:10.1097/01.chi.0000187243.17824.6c

Witt, J., VanDerHeyden, A., & Gilbertson, D. (2004). Troubleshooting behavioral interventions. A systematic process for finding and eliminating problems. *School Psychology Review*, *49*, 156–167.

*Witvliet, M., van Lier, P. A. C., & Cuijpers, P. (2009). Testing links between childhood positive peer relations and externalizing outcomes through a randomized controlled intervention study. *Journal of Consulting and Clinical Psychology*, *77*, 905–915. doi:10.1037/a0014597

Authors

ANDREA FLOWER is an assistant professor of special education at The University of Texas at Austin, 1 University Station/D5300, Austin, TX 78712; e-mail: *AndreaFlower@austin.utexas.edu*. Her primary research interests focus on the use of positive behavior support strategies to facilitate academic and behavioral change. She is also interested in teacher preparation for managing challenging behavior in school settings.

JOHN W. MCKENNA completed his doctoral degree at the University of Texas at Austin. He is an assistant professor in the Department of Special Education at St. John's University, 8000 Utopia Pkwy, Queens, NY 11439; e-mail: *mckennj1@stjohns.edu*. His primary research interests are positive behavior supports, effective instructional strategies for students with EBD and at risk. He is also interested in responsible inclusion and wrap around services.

ROMMEL L. BUNUAN is a doctoral candidate in the Department of Educational Psychology, The University of Texas at Austin, 1 University Station/D5300, Austin,

TX 78712; e-mail: *rommel.l.bunuan@gmail.com*. His research interests include multilevel modeling with a focus on mediation and meta-analysis of single-subject experimental designs studies.

COLIN S. MUETHING is a doctoral student in the Department of Educational Psychology, The University of Texas at Austin, 1 University Station/D5300, Austin, TX 78712; e-mail: *colinmuething@gmail.com*. His research interests involve the assessment and treatment of challenging behavior in individuals with developmental disabilities.

RAMON VEGA JR. is a doctoral student in the Department of Special Education, The University of Texas at Austin, 1 University Station/D5300, Austin, TX 78712; e-mail: *ramon.vega@utexas.edu*. His primary research interests are behavioral interventions for students with challenging behavior as well as examining the degree of contingency between events and behavior in a functional behavior assessment.