Perspectives on Early Childhood Psychology and Education

Volume 7
Issue 1 *Enhancing Behavioral Outcomes in Early Childhood*

Article 5

January 2023

The Effect of Token Economies on Student Behavior in the Preschool Classroom: A Meta-Analysis

Lynda B. Hayes

Brad A. Dufrene

Crystal Taylor

D. Joe Olmi

Leonard Troughton

See next page for additional authors

Follow this and additional works at: https://digitalcommons.pace.edu/perspectives

Recommended Citation

Hayes, Lynda B.; Dufrene, Brad A.; Taylor, Crystal; Olmi, D. Joe; Troughton, Leonard; Dart, Evan H.; and Weaver, Caitlyn M. (2023) "The Effect of Token Economies on Student Behavior in the Preschool Classroom: A Meta-Analysis," *Perspectives on Early Childhood Psychology and Education*: Vol. 7: Iss. 1, Article 5.
Available at: https://digitalcommons.pace.edu/perspectives/vol7/iss1/5

This Article is brought to you for free and open access by DigitalCommons@Pace. It has been accepted for inclusion in Perspectives on Early Childhood Psychology and Education by an authorized editor of DigitalCommons@Pace. For more information, please contact nmcguire@pace.edu.

The Effect of Token Economies on Student Behavior in the Preschool Classroom: A Meta-Analysis

Authors

Lynda B. Hayes, Brad A. Dufrene, Crystal Taylor, D. Joe Olmi, Leonard Troughton, Evan H. Dart, and Caitlyn M. Weaver

This article is available in Perspectives on Early Childhood Psychology and Education: https://digitalcommons.pace.edu/perspectives/vol7/iss1/5

The Effect of Token Economies on Student Behavior in the Preschool Classroom: A Meta-Analysis

Lynda B. Hayes, Brad A. Dufrene, Crystal Taylor, D. Joe Olmi, Leonard Troughton, Evan H. Dart, and Caitlyn M. Weaver

Abstract

There has been a recent push in the literature to identify and use more evidence-based practices for positive behavioral supports for challenging student behaviors in the classroom environment. Further, interest in targeting early education environments such as preschool has been growing given the persistence of behavioral difficulties in the absence of early and effective intervention (Campbell & Ewing, 1990; Kazdin, 1987; Powell et al., 2006; Stormont, 2002). Two previous meta-analyses (Maggin et al., 2011; Soares et al., 2016) provided some initial support for effectiveness of token economies with challenging student behavior; however, the inclusion of the preschool setting was limited and both studies used older versions of design standards to evaluate the quality of studies in the literature. The present study served to extend those meta-analyses by targeting preschool classrooms. Further, the current study included the most recent What Works Clearinghouse Design Standards to evaluate whether token economies meet criteria as an evidence-based practice. Ten studies were included in the final analyses. Two sets of effect sizes were calculated: Baseline-Corrected Tau and Hedge's g. An omnibus effect size showed an overall large effect; however, similar to previous meta-analyses, several methodological concerns were identified. Moderator analyses for several variables were conducted; however, no moderator analyses were significant. Limitations and future directions were discussed.

Keywords: *token economy, preschool student behavior, disruptive behavior, design standards*

The Effect of Token Economies on Student Behavior in the Preschool Classroom: A Meta-Analysis

Introduction

Researchers and educators are interested in evidence-based universal classroom management procedures for preschool classrooms. Relative to research in K-12 classrooms, far less research has been conducted testing classroom management strategies at the preschool level (Soares et al., 2016). Overall, student risk for emotional and behavior disorders (EBDs) is increasing (Pastor & Reuben, 2015). Further, preschool children's rates of EBDs are similar to rates in older children and may be higher for specific diagnoses (e.g., oppositional defiant disorder, anxiety disorders; Egger & Angold, 2006). Preschool children with EBDs may exhibit a host of symptoms, including both internalizing (e.g., withdrawal, anxiety) and externalizing (e.g., aggression, property destruction). These types of symptoms hinder children's development and success in behavioral and academic domains (Nelson et al., 2004). Further, negative outcomes such as school and social failure occur more often for children that have or are at risk for EBDs when compared to their peers. In fact, research findings indicate that over 30% of children with EBDs may drop out of high school (U.S. Department of Education, 2020), and since the 1990s, dropout rates in this category have been higher than in any other disability category.

Preschool is a critical period for identifying students who are at risk, and providing them with successful supports to increase their chances of success in academic and behavioral domains and their overall school readiness. For example, Bulotsky-Shearer et al. (2011) evaluated predictors of school readiness (e.g., early literacy, early mathematics, social-emotional competence, peer relations) and found that problem behavior (e.g., inattention, poor turn-taking skills with peers) exhibited early in the preschool academic year predicted academic outcome, motivation, attention, and

persistence with future tasks. Given these findings, researchers and preschool educators should evaluate universal classroom management systems that support preschool children's behavioral and academic development.

Token Economies

Token economies have been implemented as universal and targeted interventions, in isolation and also within larger tiered systems of support (Boerke & Reitman, 2011). Token economies have been studied for several decades and are generally shown to be effective (Doll et al., 2013). Moreover, token economies are based on fundamental principles of learning, such as positive reinforcement, and serve as the foundation for the most widely researched classroom-based interventions, such as the Good Behavior Game (Barrish et al., 1969). A benefit of the token economy is its utility in both the behavior management of an individual child or a group of children (e.g., class wide; Drabman et al., 1974; Filcheck et al., 2004; Klimas, 2007; McGoey & DuPaul, 2000; Reitman et al., 2004). Although there have been a number of variations of the token economy, the key feature is the immediate delivery of a conditioned reinforcer (e.g., token, points, sticker) after an individual (or group) exhibits a particular target behavior or class of behaviors. The token can later be exchanged for a backup reinforcer, typically from a reward menu of items pre-determined as reinforcing for the individual. The key benefit of the token economy is the ability to bridge the delay between a target behavior and the delivery of the terminal reinforcer. Bridging the delay between behavior and reinforcement is important, as delays have been shown to potentially weaken the effects of reinforcers (Boerke & Reitman, 2011; Doll et al., 2013).

Maggin et al. (2011) and Soares et al. (2016) conducted meta-analyses and design standard reviews of the token economy in the school-based literature. These meta-analyses calculated effect sizes to quantitatively synthesize the findings of studies and design

standards and evaluated the methodological rigor of studies using standards described by the What Works Clearinghouse (WWC; Kratochwill et al., 2010). Maggin et al. (2011) was the first meta-analysis conducted on token economies in the school literature that evaluated the methodological rigor of the included studies. This analysis included a total of 24 studies of the effects of token economies on student behavior. Effect sizes of the studies indicated overall improvements in student behaviors, and offered some initial support for the effectiveness of token economies implemented in the school setting on either the individual-student or class-wide level. However, the evaluation of the quality of these studies indicated several weaknesses that do not support token economies as an evidence-based practice. For example, many studies failed to meet WWC design standards (Kratochwill et al., 2010) such as insufficient demonstrations of treatment effect or three or fewer data points per phase.

Soares et al.'s (2016) results were similar to those of Maggin et al. (2011) in that token economies produced overall improvements in student behavior across the 28 included studies. In fact, approximately 25% and 68% of studies produced medium and large effect sizes, respectively. Soares et al. (2016) also evaluated the overall quality of the included studies, and results suggested the number of studies in this body of literature demonstrating acceptable standards of quality may be higher than Maggin et al. (2011); however, about 39% of included studies still demonstrated weak quality.

Overall, Maggin et al. (2011) and Soares et al. (2016) found that token economies implemented in school settings show favorable effects on student behavior in the classroom. However, there are notable limitations to both meta-analyses that warrant further investigation. First, there is a limited number of studies that included preschool populations in these meta-analyses. In fact, Maggin et al. (2011) only included K-12 in the inclusion criteria for their meta-analysis, excluding preschool children, and Soares

et al. (2016) only included 6 studies with preschool-aged children. Further, both meta-analyses utilized previous versions of WWC design standards (Kratochwill et al., 2010). WWC Version 4.1 (WWC, 2020) is an updated version including design standards that are more stringent than previous versions. Further, meta-analyses that evaluate the degree to which studies meet WWC Design Standards typically use an all-or-nothing approach. That is, studies are typically labeled as "Meets Standards," "Meets with Reservations," or "Does Not Meet" whether it fails to meet only one of the design standards or fails to meet all the standards. It may be important to parse out the degree to which a study meets each standard separately. While all standards are equally important, it may be particularly important for replication studies to know which design standards current token economy studies fail to meet. Further, it may also be the case that studies that meet a higher number of design standards yield a stronger effect size than studies that meet fewer design standards.

Although the above literature review outlined several studies that implemented variations of a token economy resulting in positive effects on student inappropriate or disruptive behavior, there are limitations of the current literature base that warrant further scientific evaluation. First, across individual and class-wide token economy studies, there are fewer studies evaluating effects for preschool-aged children compared with older students (e.g., ages 6 to 15 years; Soares et al., 2016). With the growing emphasis on early intervention strategies (Feil et al., 2016; Fox et al., 2002; Stormont, 2002), studies that evaluate viable strategies in the preschool setting are essential. Second, of the token economy strategies utilized in the preschool setting, many studies used a level system strategy and response cost (a procedure in which tokens are removed following inappropriate behavior; e.g., Filcheck et al., 2004; Reitman et al., 2004), and the effect of other strategies within this setting should be further evaluated.

Purpose of the Current Study

The current meta-analysis determined the effect size of single-case design token economy studies implemented within the preschool setting. This meta-analysis focused on studies using single subject research designs. Historically, token economy studies have used single subject research designs; and as such, limiting this meta-analysis to single subject research designs allows for a common metric for evaluating intervention effect. Additionally, this study evaluated the methodological rigor of studies included in the meta-analysis. Finally, this study included an evaluation of moderators of the effects of token economies in preschool settings. The following research questions were addressed:

1. What is the effect of token economies implemented in the preschool classroom setting on child behavior?

2. Is the effectiveness of token economies on preschool children's behavior impacted by moderator variables (e.g., number of WWC design standards met, interventionist type, primary dependent variable, design type, and presence of response cost)?

3. To what degree do token economies in preschool settings meet current design standards?

Method

Literature Search

A literature review was conducted using a multi-step process, ensuring the included articles were most appropriate to the current research questions. Two relevant, readily available databases were used: APA PsycInfo and Psychology and Behavioral Sciences Collection. Three groups of keywords were searched within the databases using Boolean Operators to target the search to more applicable studies: "preschool" or "early childhood" or "head start" or "prek" or "pre-k" AND "token economy" or "tokens" or "token" or "token system" AND "classroom."

Articles were then examined and included if they met the following inclusion criteria: 1) utilized single-case design, 2) participants were preschool-aged (2 to 5 years old), 3) setting was

the preschool classroom, 4) the study evaluated the effect of token reinforcement on student behavior, 5) article was published in a peer-reviewed journal, 6) article was available in English, and 7) publishing year was 1980 or after. The references for the articles were searched to identify any additional articles, and subsequent abstracts or full manuscripts of relevant articles were reviewed for inclusion criteria.

Article Coding

Each article was coded for four general categories: WWC Design Standards, participant characteristics, study characteristics, and interventionist characteristics. Based on WWC Design Standards 4.1 (WWC, 2020), each design standard was coded separately as "Meets Without Reservations," "Meets with Reservations," or "Does Not Meet." Two additional variables were added that computed the percentage of design standards met as well as an absolute variable (i.e., coding as "Met" required all standards to be met; coding as "Does Not Meet" required only a single standard not being met). Six separate design standard variables were coded based on WWC Version 4.1 and included the following: data availability (data must be presented visually, either in a graphical or tabular format), systematic manipulation (the experimenter must decide when and how the independent variable is manipulated), interobserver agreement (IOA; at least 20% of the data within each phase must be collected across two separate observers simultaneously and the agreement between the data must be 80% or greater), residual effects (for studies with three or more intervention types, it must be determined that there are no residual treatment effects), attempts at intervention (three attempts must be made to show a treatment effect), and meet the minimum phase length and minimum threshold of data points per phase depending on the intervention type. Although within the WWC Version 4.1 Design Standards, the phase length and minimum data points per phase is grouped into one standard, this standard was separated into two variables for the purpose of this meta-analysis.

For participant characteristics the following variables were coded: whether the study reported participant ethnicity, percentage of participants that were female, percentage of participants that were male, age range of participants, mean age of participants, special education status of participants, and socioeconomic status of participant families. Study characteristic variables included: study setting, geographic location, whether maintenance or generalization data were collected, design type, primary dependent variable and its method, and intervention components (e.g., presence of response cost, exchange schedule). Additional variables included whether the study included data on treatment integrity and social validity. Interventionist characteristics included the primary interventionist's status (e.g., teacher/staff, experimenter). Several variables were used in moderator analyses to determine whether specific variables moderate or impact the effectiveness of token economies on the behavior of preschool students. Moderator variables included: Design type, setting, components, interventionist status, percentage of WWC design standards met, overall WWC design standards, and primary dependent variable. Of note, a total of 32 variables were originally coded; however, several variables were not retained for descriptive or statistical analyses due to lack of reporting across all studies (e.g., interventionist age, interventionist years of experience); however, all original variables were coded for intercoder agreement.

Data Extraction

DigitizeIt Version 2.5 (Bormann, 2012; Rakap et al., 2016) was used to extract each numerical data point from an image of the graphs for each article to calculate effect sizes. Steps of extracting data for each article included the following: 1) taking a screen shot of each graph, 2) pasting the screenshot into the DigitizeIt software, 3) clicking on the minimum and maximum values for both the X and Y axes, and 4) clicking the center of each data point. Values for each data point were then retrieved from the software and entered into Excel for analyses. Prior to final analyses,

negative values (determined to result from extraction errors) were changed to 0.

Interrater Agreement

Agreement between the primary author and trained graduate students was calculated on several variables. Independent literature searches were conducted and discrepancies in article inclusion were discussed until 100% agreement was reached. For variable coding, the primary author trained a secondary coder on the coding scheme until 100% agreement in coding was met on a practice article. Label codes were created for the 10 articles included in the current study, label codes were randomized, and the first 3 were chosen for secondary coding (i.e., 30% of articles). Coding agreement used an extract agreement method across variables. Agreement percentage was calculated by dividing the number of variables agreed by the total number of variables and multiplied by 100. Average agreement was 84.38% across all variables (range = 0% - 100%). If agreement for a single variable fell below 80%, the raters discussed the codes until 100% agreement was reached.

The secondary coder also extracted data with Digitize It for 30% of the articles. Agreement on datum to the nearest whole number was calculated using the exact agreement method, as well as a calculation of proportional agreement in which the smaller number was divided by the larger number and multiplied by 100. Exact agreement was within an acceptable range (M = 85.28%, range = 88.79% - 98.27%). Proportional agreement was also calculated and found to also be within an acceptable range (M = 92.61%, range = 88.79% - 98.27%).

Effect Sizes

Baseline-corrected Tau (Tarlow, 2017) was utilized, which is an effect size statistic appropriate for single case design studies and incorporates both overlap of data points between phases as well as any present baseline trend. Categorical qualifiers outlined

by Vannest and Ninci (2015) are used to determine the extent to which the effect size is small (< 0.2), moderate (0.2 – 0.6), large (0.6 – 0.8), or very large (> 0.8). A free calculator available online (Tarlow, 2016) was used to calculate baseline-corrected Tau using A-B contrasts where A was a baseline phase and B was an adjacent treatment phase (Parker & Brossart, 2006). Of note, maintenance or follow up data were not included in phase contrasts for the current meta-analysis as the aim of the current analysis was on initial treatment effects. If trends in the baseline data were found, the calculator applied the baseline correction prior to calculating the final effect size. If trends in the baseline data were not found, Tau (without baseline correction) was used to calculate the final effect size.

Given the lack of consensus regarding the best effect size calculation for single case designs and to increase confidence in results, Hedge's g was also calculated for each study and across studies to produce an omnibus effect size. Interpreting Hedge's g uses the same rules of thumb as Cohen's d: 0.2 is interpreted as a small effect, 0.5 is interpreted as a medium effect, and 0.8 is interpreted as a large effect (Cohen, 1992). For analysis, means and standard deviations for each phase of each study were calculated using Microsoft Excel and entered into R (Harrer et al., 2019a; R Core Team, 2013) using the same phase contrasts as baseline-corrected Tau. Within R, the dmetar package was utilized (Harrer et al., 2019b). Due to differences in sampling across studies, a random effects model was used to calculate the omnibus effect of token economies on preschool students' behavior.

Results

Literature Search

The initial phase of the literature search with the included Boolean operators yielded 42 articles across both the APA PsycInfo and Psychology and Behavioral Sciences databases. Initially, abstracts were reviewed and studies were excluded if they failed

to meet any of the 6 inclusion criteria. For the remaining articles, the manuscripts were reviewed in full to determine if each study met inclusion criteria. Based on these inclusion criteria, 10 articles were retained for the meta-analysis. The author included one additional study following the ancestral search (i.e., Wolfe et al., 1983); however, the study was ultimately excluded due to graphical representation of the data that could not be extracted using the current methods. In total, 10 articles were determined to meet inclusion criteria for the current meta-analysis.

Descriptive Statistics

WWC Design Standards

None of the included studies fully met WWC Version 4.1 (WWC, 2020) design standards, as each study failed to "Meet without Reservations" on at least one design standard variable and only 20% of articles met all criteria with reservations (see Table 1).

Table 1

WWC Design Standards

	DS1	DS2	DS3	DS4	DS5	DS6	Percentage of Standards Met
Tiano et al. (2005)	MS	MS	MS	NA	MS	DNM	80%
McGoey & DuPaul (2000)	MS	MS	DNM	NA	MS	MWR	60% (80%*)
Filcheck et al., 2004	MS	MS	DNM	NA	DNM	MWR	40% (60%*)
Plavnick et al., 2010	MS	MS	DNM	NA	DNM	DNM	40%
Reitman et al., 2004	MS	MS	MS	MS	MS	MWR	83.33% (100%*)
Sran & Borrero, 2010	MS	MS	DNM	NA	MS	MS	80%
Swiezy et al., 1993	MS	MS	DNM	NA	MS	MWR	60% (80%*)
Miller et al., 1981	MS	MS	DNM	DNM	MS	MS	66.67%
Conyers et al., 2004	MS	MS	MS	MS	MS	MWR	83.33% (100%*)
Conyers et al., 2003	MS	MS	DNM	NA	MS	DNM	60%

Note. *DS1 = Data availability, DS2 = Systematic manipulation, DS3 = Interobserver agreement, DS4 = Residual effects, DS5 = Attempts at intervention effect, DS 6 = Data points per phase, MS = Meets standard without reservation, MWR = Meets standard with reservation, DNM = Does not meet standard, NA = Not applicable. An asterisk (*) indicates percentages of standards met without or with reservations.*

All of the included studies met design standards for data availability and systematic manipulation. Only 30% of the studies met the design standard regarding IOA. The design standard related to residual effects was met by 66.67% of studies of which this design standard was applicable (i.e., 2 of 3 studies). Eighty percent of the studies met the attempts at intervention effects design standard. Twenty percent of the studies met the design standards for minimum data points per phase without reservations, and 50% of the studies met the design standards for minimum data points per phase with reservations.

Participant, Interventionist, and Study Characteristics

Overall, there were limited data provided for participant and interventionist characteristics across studies. For example, 70% of studies failed to report race or ethnicity data for participants. Studies that did report these data show that most participants were white or Caucasian and male (64.74%). Although all the included studies took place in a preschool classroom setting, location types varied across the set of studies: 60% took place in a regular, public preschool classroom while 20%, 10%, and 10% of studies took place in Head Start classrooms, special education classrooms, and parochial preschool classrooms, respectively. Interventionists in 60% of the studies were preschool classroom teachers or staff and the remaining were experimenters. Of note, one study did not report status of the interventionist (Conyers et al., 2003).

The majority (40%) of studies utilized an alternating treatments or multielement design. The remaining studies used a reversal (20%), withdrawal (20%), or multiple baseline (20%) design. Each study's primary dependent variable was coded into two general categories: inappropriate student behavior or appropriate student behavior with most studies using inappropriate student behavior as the primary dependent variable. Examples of inappropriate student behavior included off-task behavior and breaking classroom rules (e.g., keep hands to self). Examples of appropriate student behavior included appropriate sitting behavior, responding to the target task,

and appropriate rest-time behavior. Half of the included studies included a response cost. Of those studies, the response cost procedure was either incorporated within the components of the token economy (60%) or directly compared to token reinforcement and response cost (40%). The exchange rate of tokens also varied across the included studies: 50% exchanged tokens once daily, 30% multiple times per day, and 20% failing to report the exchange rate.

Treatment integrity data were reported in five studies. Tiano et al. (2005) reported treatment integrity was above 85% and no retraining was necessary throughout the study. McGoey and DuPaul (2000) reported treatment integrity remained at 100% across all phases of the study; however, the researchers only checked treatment integrity once per week. Across all phases in Filcheck et al. (2004), average treatment integrity was reported to be 67.8% and a total of seven retrainings were required across the duration of the study. Plavnick et al. (2010) reported an average treatment integrity of 84% across the teacher participants. Finally, although Swiezy et al. (1993) reported they collected data on treatment integrity, the authors did not provide the data within the article.

Social validity data were reported in 4 studies (Filcheck et al., 2004; McGoey & DuPaul, 2000; Reitman et al, 2004; Tiano et al., 2005). However, two of those studies failed to report specific outcomes (Filcheck et al., 2004; Tiano et al., 2005. Both studies that did report outcomes used teacher-rated treatment acceptability as the measure of social validity based on the Intervention Rating Profile-15 (IRP-15; Martens et al., 1985). The IRP-15 consists of 15 items rated on a Likert scale ranging from 1 (not acceptable) to 6 (very acceptable). Total scores on the IRP-15 range from 15 to 90 and higher scores represent higher acceptability. McGoey and DuPaul (2000) reported a per-item average rating of 5.1 representing high acceptability. Reitman et al. (2004) reported varied acceptability across all 3 participants: poor (IRP-15 = 20), moderate (IRP-15 = 61), and high (IRP-15 = 83).

Forty percent of the included studies reported a maintenance or follow up phase. Of those studies, one study reported the maintenance phase began immediately after the final intervention

phase (Miller et al., 1981), one study reported the maintenance phase began within 1 month of the final intervention phase (McGoey & DuPaul, 2000), and two studies reported the maintenance phase began at or more than one month after the final intervention phase (Filcheck et al., 2004; Tiano et al. 2005). Swiezy et al. (1993) evaluated the degree to which their treatment effects in the classroom generalized to the school playground and was the only study that reported generalization data.

Effect Size Calculations

Baseline-Corrected Tau

A total of 63 phase contrasts across studies were analyzed to calculate Baseline-Corrected Tau effect sizes. No baseline corrections were necessary and the final effect size was calculated using Tau (without baseline correction). Overall, effect sizes across studies ranged from 0 to 0.745 with a mean of 0.499. See Table 2 for Baseline-Corrected Tau effect sizes across phase contrast within each study.

Table 2

Baseline-Corrected Tau Across Studies

Study	Participant	Phase Contrast	Baseline-Corrected Tau	Effect Size
Tiano et al. (2005)	Ruby	BL1-RC	0.745	Large
		BL2-TE^{R}	0.215	Moderate
	Damon	BL1-RC	0.537	Moderate
		BL2-TE^{R}	0.000	Small
	Mitch	BL1-RC	0.566	Moderate
		BL2-TE^{R}	0.336	Moderate
McGoey & DuPaul (2000)	Derek	BL1-TE1	0.728	Large
		BL2-RC1	0.252	Moderate
		BL3-TE2	0.775	Large
		BL4-RC2	0.378	Moderate
	Douglas	BL1-TE1	0.542	Moderate
		BL2-RC1	0.478	Moderate
		BL3-TE2	0.775	Large
		BL4-RC2	0.258	Moderate
	Monica	BL1-RC1	0.726	Large
		BL2-TE1	0.630	Large
		BL3-RC2	0.258	Moderate
		BL4-TE2	0.756	Large
	Rebecca	BL1-RC1	0.189	Small
		BL2-TE1	0.629	Large
		BL3-RC2	0.602	Large
		BL4-TE2	0.775	Large
Filcheck et al. (2004)	Classwide	BL1-TE^{R}	0.411	Moderate
		BL2-CDI	0.463	Moderate
		BL2-PDI	0.693	Large
Plavnick et al. (2010)	Toby	BL-TE	0.399	Moderate
	Kendra	BL-TE	0.213	Moderate
Conyers et al. (2004)	Classwide	BL1-RC1	0.622	Large
		BL2-TE	0.639	Large
		BL2-RC2	0.510	Moderate
Conyers et al. (2003)	Classwide	BL1-TE1	0.603	Large
		BL2-TE2	0.366	Moderate
Reitman et al. (2004)	Simon	BL1-GR1^{R}	0.539	Moderate
		BL1-IN1^{R}	0.346	Moderate
		BL2-GR2^{R}	0.679	Large
		BL2-IN2^{R}	0.680	Large
	Xavier	BL1-GR1^{R}	0.396	Moderate
		BL1-IN1^{R}	0.693	Large
		BL2-GR2^{R}	0.658	Large
		BL2-IN2^{R}	0.756	Large
	Tom	BL1-GR1^{R}	0.587	Moderate
		BL1-IN1^{R}	0.702	Large
		BL2-GR2^{R}	0.648	Large
		BL2-IN2^{R}	0.770	Large
Sran & Borrero (2010)	Dylan	BL-NO	0.065	Small
		BL-SI	0.287	Moderate
		BL-VA	0.348	Moderate
	Mira	BL-NO	0.367	Moderate
		BL-SI	0.472	Moderate
		BL-VA	0.472	Moderate
	Milo	BL-NO	0.147	Small
		BL-SI	0.219	Moderate
		BL-VA	0.219	Moderate
	Luke	BL-NO	0.339	Moderate
		BL-SI	0.139	Small
		BL-VA	0.261	Moderate
Swiezy et al. (1993)	Pair A	BL1-TE1	0.518	Moderate
		BL2-TE2	0.518	Moderate
	Pair B	BL1-TE1	0.724	Large
		BL2-TE2	0.655	Large
Miller et al. (1981)	Classwide	BL1-TE1	0.716	Large
		BL2-TE2	0.730	Large
		BL2-TE3	0.745	Large

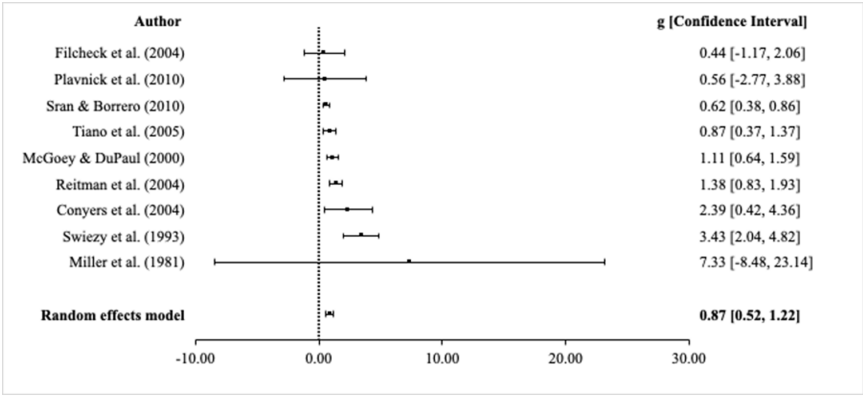
Note. *BL = Baseline, TE = Token Economy, RC = Response Cost, CDI = Child Directed Interaction, PDI = Parent Directed Interaction, GR = Group Token Economy, IN = Individual Token Economy, NO = No Choice, SI = Single Choice, VA = Varied Choice. Superscript R denotes token economies that also included a response cost component.*

Hedge's g

Hedge's g was computed for each of the 10 included studies (see Table 3). The majority of studies produced a large effect size based on the rule of thumb (i.e., 0.8 threshold; Cohen, 1992). Filcheck et al. (2004)'s effect size was small (0.4425). Plavnick et al. (2010) and Sran and Borrero (2010) reported medium effect sizes. See Table 4 for Hedge's g effect sizes, confidence intervals, and standard errors for all studies. See Figure 1 for a forest plot of effect sizes for each study.

Figure 1

Forest Plot of Effect Sizes by Study



Note. *Conyers et al. (2003) was removed from the final forest plot due to inability to interpret the forest plot with it included (due to its wide confidence interval [-66.16 to 81.68]).*

Hedge's g was also calculated across all included studies to produce an omnibus effect size. The omnibus effect size using Hedge's g was 0.8704, $p = 0.003$ and is considered a large effect size. The included studies were analyzed to determine whether there were outliers present. The find.outliers function within the dmetar package detected an outlier based on a significant test of heterogeneity ($p = 0.0033$). The outlier (Swiezy et al., 1993) was removed from analysis and the test of heterogeneity was not significant ($p = 0.1762$) for the final omnibus effect size calculation.

With the outlier removed, Hedge's g was 0.8257, $p < 0.0001$ and is also considered a large effect size (Cohen, 1992).

Table 3

Effect Size by Study

Study	Number of Contrasts	Hedge's g	Confidence Intervals		Standard Error
			Lower	Upper	
Tiano et al. (2005)	6	0.8694[L]	0.3686	1.3701	0.25548469
McGoey & DuPaul (2000)	16	1.1138[L]	0.6352	1.5924	0.24418367
Filcheck et al. (2004)	3	0.4425[S]	-1.1727	2.0576	0.82405612
Plavnick et al. (2010)	2	0.5574[M]	-2.7681	3.883	1.69670918
Conyers et al. (2004)	3	2.3889[L]	0.4186	4.3592	1.0052551
Conyers et al. (2003)	2	7.7557[L]	-66.1653	81.6766	37.7147704
Reitman et al. (2004)	12	1.3796[L]	0.8318	1.9274	0.2794898
Sran & Borrero (2010)	12	0.6208[M]	0.38	0.8615	0.12283163
Swiezy et al. (1993)	4	3.4279[L]	2.0383	4.8174	0.70895408
Miller et al. (1981)	2	7.3282[L]	-8.4839	23.1403	8.06739796

Note. *The superscript S denotes a small effect, the superscript M denotes a medium effect, and superscript L denotes a large effect.*

Moderator Analyses

Moderator analyses were conducted for seven variables to determine their effects on the impact of token economies on preschool student behavior: Design Type, Setting, Inclusion of Response Cost, Interventionist Status, Number of WWC Standards Met, Overall WWC, and Primary Dependent Variable. Although medium to large effects were found, none of the analyses produced significant results on student behavior outcomes (see Table 4).

Table 4
Effect Sizes for Moderator Variables

Moderator	Category	K (studies)	Hedge's g	95% Confidence Interval	
				Lower	Higher
Design Type	Withdrawal/Reversal	4	0.9729^{L}	0.6648	1.281
	Alternating Treatments	2	0.7652^{M}	0.1329	1.3976
	Multiple Baseline	4	2.4066^{L}	-15.0553	19.8684
Setting	Head Start	2	1.1119^{L}	-2.1255	4.3493
	Public Preschool	6	0.7361^{M}	0.4118	1.0604
	Special Education Preschool	1	0.5574^{M}	-2.7681	3.883
	Church Preschool	1	3.4279^{L}	2.0383	4.8174
Components	With Response Cost	5	1.1342^{L}	0.5781	1.6904
	Without Response Cost	5	0.734^{M}	0.0345	1.3724
Interventionist Status	Teacher	6	1.0832^{L}	0.7852	1.3813
	Experimenter	3	1.9733^{L}	-1.7495	5.6961
Percent of WWC Standards Met	66.67%	3	0.8576^{L}	-3.3891	5.1042
	83.33%	5	1.2291^{L}	-0.1725	2.6308
	100%	2	1.452^{L}	-1.8577	4.7617
Overall WWC	Met	2	1.452^{L}	-1.8577	4.7617
	Does Not Meet	8	0.7919^{M}	0.4118	1.172
Primary DV	Appropriate	4	0.7034^{M}	-0.1811	1.5878
	Inappropriate	6	1.1125^{L}	0.7732	1.4517

Note. *The superscript S denotes a small effect, the superscript M denotes a medium effect, and superscript L denotes a large effect.*

Discussion

Although two recent meta-analyses were conducted evaluating the effect of token economies in classrooms, (Maggin et al., 2011; Soares et al., 2016), the current meta-analysis attempted to expand on those results by targeting the preschool setting and including the latest WWC Version 4.1 Design Standards (WWC, 2020). Similar to the results of Maggin et al. (2011) and Soares et al. (2016), results of the current meta-analysis showed that token economies generally produce favorable and large effects on increasing appropriate behavior and decreasing inappropriate behavior in the preschool classrooms. In the Maggin et al. (2011) and Soares et al. (2016) meta-analyses, the overall effect was large.

However, the preschool setting was not evaluated in Maggin et al. (2011). Soares et al. (2016) did include the preschool setting, and their moderator analysis showed a statistically lower effect size for ages 3 to 5 compared to 6 to 15. The number of articles included in the current meta-analysis represented approximately a 67% increase from the number of preschool articles included in Soares et al. (2016). There was some considerable overlap in the preschool articles included in both studies; specifically, five articles were included in the current meta-analysis and Soares et al. (2016). The inclusion criteria used by Soares and colleagues was limited to public preschool classrooms whereas the current meta-analysis expanded this to other settings (e.g., special education classrooms, parochial classroom); thus, the results of the current meta-analysis may be more generalizable than the results of Soares et al. (2016).

Maggin et al. (2011) and Soares et al. (2016) also evaluated methodological rigor of token economy studies; however, both studies used older WWC standards (Kratochwill et al., 2010). The current meta-analysis reviewed studies utilizing the most recent design standards (WWC, 2020), which are more rigorous than previous WWC standards. Soares et al. (2016) found that token economy studies in preschool settings did not meet design standards; in fact, 50% of the preschool studies included in the meta-analysis were weak (i.e., did not meet standards). Results from this study are consistent with those findings. None of the 10 studies included in this meta-analysis met design standards without reservations based on the most recent standards (WWC, 2020). Moreover, eight studies did not meet standards with reservations. These results indicate that researchers and practitioners must be cautious with regard to interpreting findings from this meta-analysis and from individual studies that have tested token economies in preschool classrooms. Poor research design and execution undermines internal and external validity. For example, if a single case design study includes less than five data points per phase and IOA data for the dependent measures were not adequately

sampled, then researchers and practitioners cannot be confident that changes in behavior are due to the intervention. It may be that changes in behavior are due to instrumentation shift or an unreliable, inadequate sample of behavior. Similarly, if treatment integrity data are not provided, then changes in behavior cannot be attributed to the independent variable. Therefore, future research testing token economies in preschool classrooms must be designed and executed with more rigorous designs and procedures.

This study also conducted moderator analyses of several variables, and results indicated no significant moderators of token economy effects. However, it is important to note that this meta-analysis only included 10 studies; results of the moderator analyses should be interpreted with caution given that the inclusion of fewer studies may significantly affect the statistical power necessary to detect differences between groups (Borenstein et al., 2009). Relatively few token economy studies have been conducted in preschool settings. As more studies accumulate, another meta-analysis may be conducted and moderator analyses may yield important moderators of token economy effects in preschool classrooms.

Limitations

Several limitations of the current meta-analysis should be considered when interpreting its results. First, the initial literature search used only two databases relevant to the social and behavioral sciences. It may be the case that expanding the search to other databases would have yielded a higher number of articles. However, an ancestral search was used to include articles not otherwise available in the initial search. Therefore, this meta-analysis may adequately sample published research testing token economies in preschool classrooms. Relatedly, a second limitation includes the limited number of total articles included in the current meta-analysis. Although it has been suggested that only two studies are needed to conduct a meta-analysis (Valentine et al., 2010) and

at least five are needed for sufficient power (Jackson & Turner, 2017), it is likely that overall conclusions of the effectiveness of token economies within the preschool classroom will change as more studies are included in future analyses and statistical power is increased. Further, it may be the case that different sets of inclusion criteria would yield a higher number of articles to include. In this meta-analysis, for example, the author only included articles that were published in peer-reviewed journals, which may be subject to publication bias (i.e., favoring publication of studies with stronger effects; Tincani & Travers, 2019). Future meta-analyses should include grey literature to increase the score of the analysis as well as increase power. Third, the author coded the dependent variables into two general categories (appropriate and inappropriate student behavior). However, specific definitions of behaviors differed across the included studies. It may be the case that token economies have a different effect on different types of student behaviors (e.g., more disruptive externalizing behaviors such as tantrumming versus more passive behaviors such as off-task). In addition, token economies have also been evaluated to improve outcomes other than student appropriate or inappropriate behaviors (e.g., academic achievement; Ayllon et al., 1972) and a meta-analysis that includes a greater variety of outcome variables may produce different effects. Finally, some studies included intervention components outside of the standard procedures of token economies, and the degree to which their presence altered the effectiveness of treatment is unknown. For example, Sran and Borrero (2010) included different variations of token economy exchanges to include a no choice condition (i.e., children exchanged tokens for only one reinforcer), single choice condition (i.e., children exchanged tokens for one of five identical reinforcers), and varied choice condition (i.e., children exchanged tokens for one of five different reinforcers).

In addition to the limitations of the current meta-analysis, limitations of the included studies should also be noted. The

majority of studies did not report data for a number of different areas, including specific treatment components, participant characteristics, and interventionist characteristics. Absence of these data limits the extent to which future researchers can attempt to replicate these studies, as well as the degree to which the studies' findings can translate from sample to population. Many studies also did not report sufficient data related to treatment integrity and social validity. Measuring and reporting treatment integrity data are crucial in regard to internal validity. If the degree to which treatment was implemented with integrity is unknown, treatment outcomes cannot be properly assessed or correlated with the treatment. Further, information regarding treatment integrity is important for external validity and the extent to which treatment may be implemented in real-world settings. Relatedly, measures or procedures to calculate treatment integrity and social validity varied across the studies that included those data. Finally, maintenance and generalization data were not collected for most studies; thus, it is unknown if treatment effects maintained over time and generalized to other settings.

Future Directions

Although this meta-analysis yielded results in favor of the overall effectiveness token economies have on children's behavior in the preschool classroom, future studies should attend to aforementioned limitations. In particular, researchers should include treatment integrity data due to its importance to internal and external validity. Researchers should provide more information in regard to interventionist characteristics, since treatment integrity may vary based on professional background and training (e.g., researchers, teacher assistants), and variations in treatment integrity may impact treatment outcomes. Moreover, IOA data are needed to strengthen the rigor of the research, and thus increase the believability of findings. Overall, major methodological changes are needed for future studies, including meeting WWC Version 4.1 Design Standards (WWC, 2020), inclusion of treatment integrity data, and inclusion of social validity

data to measure the degree to which token economies produce meaningful and sustainable changes to the classroom environment.

References

Note: Asterisk (*) denotes articles included in the current meta-analysis

Ayllon, T. & Kelly, K. (1972). Effects of reinforcement on standardized test performance. *Journal of Applied Behavior Analysis, 5*(4), 477-484. https://doi.org/10.1901/jaba.1972.5-477

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*(2), 119-124. https://doi.org/10.1901/jaba.1969.2-119

Boerke, K. W., & Reitman, D. (2011). Token economies. In W. W. Fisher, C. C. Piazza, & H. S. Roane (Eds.), Handbook of applied behavior analysis (pp. 370-382). The Guilford Press.

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

Bormann, (2012) DigitizeIt (version 2.0) http://www.digitizeit.de/

Bulotsky-Shearer, R. J., Fernandez, V., Dominguez, X., & Rouse, H. L. (2011). Behavior problems in learning activities and social interactions in Head Start classrooms and early reading, mathematics, and approaches to learning. *School Psychology Review, 40,* 39-56. https://doi.org/10.1080.02796015.2011.12087727

Campbell, S. B. & Ewing, L. J. (1990). Follow-up of hard-to-manage preschoolers: Adjustment at age 9 and predictors of continuing symptoms. *Journal of Child Psychology and Psychiatry, 6,* 871-889. https://doi.org/10.1111/j.1469-7610.1990.tb00831.x

Cohen, J. (1992). Statistical power analysis. Current Directions in Psychological Science, 1(3), 98-101. https://doi.org/10.1111/1467-8721.ep10768783

*Conyers, C., Miltenberger, R., Maki, A., Barenz, R., Jurgens, M., Sailer, A., Haugen, M., & Kopp, B. (2004). A comparison of response cost and differential reinforcement of other behavior to reduce disruptive behavior in a preschool classroom. *Journal of Applied Behavior Analysis, 37*(3), 411-415. https://doi.org/10.1901/jaba.2004.37-411

*Conyers, C., Miltenberger, R., Romaniuk, C., Kopp, B., & Himle, M. (2003). Evaluation of DRO schedules to reduce disruptive behavior in a preschool classroom. *Child & Family Behavior Therapy, 25*(3). 1-6. https://doi.org/10.1300/J019v25n03_01

Drabman, R. S. & Lahey, B. B. (1974). Feedback in classroom behavior modification: Effects on the target and her classmates. *Journal of Applied Behavior Analysis, 7*(4), 591-598. https://doi.org/10.1901/jaba.1974.7-591

Doll, C., McLaughlin, T. F., Barretto, A. (2013). The token economy: A recent review and evaluation. *International Journal of Basic and Applied Science, 02*(01), 131-149.

Egger, H. L. & Angold, A. (2006). Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry, 47*(3/4), 313-337. https://doi.org/10.1111/j.1469-7610.2006.01618.x

Feil, E. G., Small, J. W., Seeley, J. R., Walker, H. M., Golly, A., Frey, A., & Forness, S. R. (2016). Early intervention for preschoolers at risk for Attention-Deficit/ Hyperactivity Disorder: Preschool First Step to Success. *Behavioral Disorders, 41*(2), 95–106. https://doi.org/10.17988/0198-7429-41.2.95

*Filcheck, H. A., McNeil, C. B., Greco, L. A., & Bernard, R. S. (2004). Using a whole-class token economy and coaching of teacher skills in a preschool classroom to manage disruptive behavior. *Psychology in the Schools, 41*(3), 351-361. https://doi.org/10.1002/pits.10168

Fox, L., Dunlap, G., Cushing, L. (2002). Early intervention, positive behavior support, and transition to school. *Journal of Emotional and Behavioral Disorders, 10*(3), 149-157. https://doi.org/10.1177/10634266020100030301

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019a). Doing Meta-Analysis in R: A Hands-on Guide. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Harrer, M., Cuijpers, P., Furukawa, T. & Ebert, D. D. (2019b). dmetar: Companion R package for the guide 'Doing Meta-Analysis in R'. R package version 0.0.9000. http://dmetar.protectlab.org/

Jackson, D. & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods, 8,* 290-302. https://doi.org/10.1002/jrsm.1240

Kazdin, A. E. (1987). Treatment of antisocial behavior in children: Current status and future directions. *Psychological Bulletin, 102*(2), 187-203. https://doi.org/10.1037/0033-2909.102.2.187

Klimas, A. & McLaughlin, T. F. (2007). The effects of token economy system to improve social and academic behavior with a rural primary and child with disabilities. *International Journal of Special Education, 22*(3).

Kratochwill, T. R., Hitchock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D.M. & Shadish, W.R. (2010). *Single-Case Designs Technical Documentation.* Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf .

Maggin, D., M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology, 49*, 529-554. https://doi.org/10.1016/j.jsp.2011.05.001

Martens, B. K., Witt, J. C., Elliott, S. N., & Darveaux, D. X. (1985). Teacher judgments concerning the acceptability of school-based interventions. *Professional Psychology: Research and Practice, 16*(2), 191-198. https://doi.org/10.1037/0735-7028.16.2.191

*McGoey, K. E. & DuPaul, G. J. (2000). Token reinforcement and response cost procedures: Reducing the disruptive behavior of preschool children with attention-deficit/ hyperactivity disorder. *School Psychology Quarterly, 15*(3), 330-343. https://doi.org/10.1037/h0088790

*Miller, M. A., McCullough, C. S., & Ulman, J. D. (1981). Carryover effects of multielement manipulations: Enhancement of preschoolers' appropriate rest-time behavior. *Educational Psychology: An International Journal of Experimental Educational Psychology, 1*(4), 341-346. https://doi.org/10.1080/0144341810010405

Nelson, J. G., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K-12 students with emotional and behavioral disorders. *Exceptional Children, 71,* 59-73. https://doi.org/10.1177/001440290407100104

Parker, R. I. & Brossart, D. F. (2006). Phase contrasts for multiphase single case intervention designs. *School Psychology Quarterly, 21*(1), 46-61. https://doi.org/10.1521/scpq.2006.21.1.46

Pastor, P. N. & Reuben, C. A. (2015). Trends in parent-reported emotional and behavioral problems among children using special education services. *Psychiatric Services, 66*(6), 656-659. https://doi.org/10.1176/appi.ps.201400254

*Plavnick, J. B., Ferreri, S. J., & Maupin, A. N. (2010). The effects of self-monitoring on the procedural integrity of a behavioral intervention for young children with developmental disabilities. *Journal of Applied Behavior Analysis, 43*(2), 315-320. https://doi.org/jaba.2010.43-315

Powell, D., Dunlap, G., & Fox, L. (2006). Prevention and intervention for the challenging behaviors of toddlers and preschoolers. *Infants & Young Children, 19*, 25-35. https://doi.org/10.1097/00001163-200601000-00004

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org/

Rakap, S., Rakap, S., Evran, D., & Cig, O. (2016). Comparative evaluation of the reliability and validity of three data extraction programs: UnGraph, GraphClick, and DigitizeIt. Computers in Human Behavior, 55, 159–166. https://doi.org/10.1016/j.chb.2015.09.008

*Reitman, D., Murphy, M. A., Hup, S. D. A., & O'Callaghan, P. M. (2004). Behavior change and perceptions of change: Evaluating the effectiveness of a token economy. *Child & Family Behavior Therapy, 26*(2), 17-36. https://doi.org/10.1300/J019v26n02_02

Soares, D. A., Harrison, J. D., Vannest, K., & McClelland, S. S. (2016). Effect size for token economy use in contemporary classroom settings: A meta-analysis and moderator analysis of single case research. *School Psychology Review, 45*(4), 379-399. https://doi.org/10.17105/SPR45-4.379-399

*Sran, S. K. & Borreo, J. C. (2010). Assessing the value of choice in a token economy. *Journal of Applied Behavior Analysis, 43*(3), 553-557. https://doi.org/10.1901/jaba.2010.43-553

Stormont, M. (2002). Externalizing behavior problems in young children: Contributing factors and early intervention. *Psychology in the Schools, 39*(2), 127-138. doi: 10.1002/pits.10025

*Swiezy, N. B., Matson, J. L., & Box, P. (1993). The good behavior game. *Child & Family Behavior Therapy, 14*(3), 21-32. https://doi.org/10.1300/J019v14n03_02

Tarlow, K. R. (2016). Baseline corrected tau calculator. http://ktarlow.com/stats/tau/

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected tau. *Behavior Modification, 41*(4), 427-467. https://doi.org/10.1177/0145445516676750

*Tiano, J. D., Fortson, B. L., McNeil, C. B., & Humphreys, L. A. (2005). Managing classroom behavior of Head Start children using response cost and token economy procedures. *Journal of Early and Intensive Behavior Intervention, 2*(1). 28-39. https://doi.org/10.1037/h0100298

Tincani, M. & Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science, 42*(1), 59-75. https://doi.org/10.1007/s40614-019-00191-5

U.S. Department of Education. (2020). *Forty-second annual report to Congress on the implementation of the Individuals with Disabilities Education Act*. Author.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*(2), 215-247. https://doi.org/10.3102/1076998609346961

Vannest, K. J. & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. Journal of Counseling & Development, 93(4), 403.411. https://doi.org/10.1002/jcad.12038

What Works Clearinghouse. (2020). *What Works Clearinghouse Standards Handbook, Version 4.1.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Wolfe, V. V., Boyd, L. A., & Wolfe, D. A. (1983). Teaching cooperative play to behavior-problem preschool children. *Education and Treatment of Children, 6*(1), 1-9.