Summer 8-1-2021

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

Lynda B. Hayes

THE EFFECT OF TOKEN ECONOMIES ON STUDENT BEHAVIOR IN THE PRESCHOOL CLASSROOM: A META-ANALYSIS

by

Lynda B. Hayes

A Dissertation
Submitted to the Graduate School,
the College of Education and Human Sciences
and the School of Psychology
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved by:

Dr. Brad A. Dufrene, Committee Chair
Dr. D. Joe Olmi
Dr. Evan Dart
Dr. Leonard Troughton

August 2021

Published by the Graduate School

THE UNIVERSITY OF SOUTHERN MISSISSIPPI®

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, Dunlap, Fox, 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 previous 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 or not 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.

ACKNOWLEDGMENTS

I would first like to thank my committee chair, Mister B.A.D. himself, Dr. Brad A. Dufrene, who provided the encouragement, balanced with humor and tough-love accountability, that was necessary to see me through this program, my practicum position, and this project. I offer my gratitude to my committee members, Drs. Joe Olmi, Evan Dart, and Leonard Troughton for offering their support and feedback throughout this process. I would also like to extend my sincere appreciation to Dr. Allison Grennan and other supportive supervisors at Munroe-Meyer Institute, as your encouragement, flexibility, and grace was integral to the completion and success of this project during internship. I would also like to thank my fellow student and intern colleagues who provided an extensive amount of their time and energy to me throughout this program and this project: Rob Derieux, Caitlyn Weaver, Taylor Altenberger, Meleah Ackley, Andrew Rozsa, III, Parker Lundy, Ashleigh Eaves, Katie Lachance (and Sage), Kaitlyn Young, and Janet Schwartz-Micheaux.

DEDICATION

This work is dedicated to the lives and memories of my son, Eli Carson Hayes-McCloskey, and my father, Stephen Patrick Hayes. Eli, it was your life that inspired me to reach for this degree. Dad, it was your support that helped me grab it. May your energies continue to guide me through life and continue reaching for goals. Ever upwards and onwards.

Humans are social creatures by nature. Our ancestors' chances of survival were significantly impacted by their network. So, too, was mine. I have suffered great losses throughout my lifetime and many hurdles that interrupted this seventeen-year journey, but that has been balanced by the incredible people that have surrounded me and continued to believe in me through the (many) years and patiently (so, so patiently) supported me through this process: my sisters and their beautiful families: Carolyn, Dominic, Mckenzie, Victoria, and Xavier; Cindy, Justin, Cameron, Collin, and Calli. As well as Brian and Terri Cleary, Daniel Hayes, Haley Moon (of my life), Timothy McCloskey (and family), Christina Lehman (and family), Erin Lavender-Stott, Logan Hoppe, Angie Lee, Dr. Carole Van Camp and many more (naming you all would surely double the length of this document). I am humbled by and grateful for the vast network of love and support I have been privileged to build over the years.

We. Did. It.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I - INTRODUCTION

In recent years, there has been an increased interest to add to the evidence-based literature in the area of positive behavioral supports for students who exhibit challenging problem behaviors in the classroom. Among these students are those who have or are at-risk for emotional and behavioral disorders (EBDs). Students with EBDs may exhibit a host of symptoms, including both internalizing (e.g., withdrawal, anxiety) and externalizing (e.g., aggression, property destruction) symptoms. These types of symptoms hinder student development and success in both the behavioral and academic domains (Nelson et al., 2004). Further, negative outcomes like school and social failures occur more often for students that have or are at-risk for EBDs when compared to their peers. In fact, data indicate that over 30% of students 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 more than in any other disability category.

Behavioral problems that present early in life have been shown to persist throughout one's lifetime in the absence of early and effective intervention (Campbell & Ewing, 1990; Kazdin, 1987; Powell et al., 2006; Stormont, 2002); thus, there has been a particular growing interest in the development and evaluation of intervention strategies during early education (e.g., preschool), especially given that positive teacher-student relationships may ameliorate some negative outcomes associated with early onset behavioral problems (Sabol & Pianta, 2012; Silver et al., 2005). Preschool-aged years are critical for identifying students who are at risk and providing them with successful supports to increase their chances of success in both the academic and behavioral domains and their overall school readiness. For example, in a recent study evaluating

predictors of school readiness, it was 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 (Bulotsky-Shearer et al., 2011).

Behavioral Interventions in Preschool

Positive Behavior Intervention and Supports

A Multi-Tiered System of Support (MTSS; McIntosh & Goodman, 2016) is one that provides effective supports for the educational success of students across both the academic and behavioral domains. Within the behavioral domain, one MTSS approach to reduce the occurrence of students' problem behaviors and increase their appropriate and adaptive behaviors in the classroom is Positive Behavior Intervention and Supports (PBIS; Carr et al., 2002; Office of Special Education Technical Assistance Center on Positive Behavioral Intervention & Supports, 2015; Sugai & Horner, 2006). The number of schools that have reported PBIS implementation has increased from approximately 14,000 in 2010 to an estimated 23,000 in 2017 (Horner et al., 2017; Sugai & Horner, 2014).

PBIS tiers of support. Support delivery within PBIS is implemented across three tiers with the overall aim to prevent or decrease student problem behaviors. Tier 1 of PBIS is the primary, or universal tier, and is implemented on a school-wide basis and its support strategies contact every student within the school system. Tier 1 supports include systems that are designed to prevent students' problem behaviors. School-wide systems may include universal screening, school-wide behavioral expectations across all settings, and consistent training and implementation of behavior management strategies across all

staff (Horner et al., 2010). Within Tier 1, class-wide behavioral management strategies are often and may include clearly communicated expectations, behavioral skills training for expected behaviors, behavior-specific praise, and corrective teaching interactions. Additionally, group contingency interventions may be utilized, such as the Good Behavior Game (Barrish et al., 1969; Tingstrom et al., 2006) and class-wide token economies (Filcheck et al., 2004). The secondary tier, or Tier 2, includes more intense level of supports for students that are considered non-responders to the primary tier. Tier 2 supports are designed to be resource efficient and prevent emerging student difficulties from worsening such that intensive intervention is required. Tier 2 supports may include small group social skills groups or interventions that are implemented in a standardized fashion (e.g., Check-in/Check-out; LaBrot et al., 2016). Additionally, students in Tier 2 receive progress monitoring (e.g., daily behavior report card) to gauge their response to supports (Chafouleas et al., 2006). Students whose behavioral data suggest they are not responsive to secondary level of supports may then be referred for the Tier 3 intervention. Within Tier 3, supports are individualized, and interventions are more intense than lower-level tiers. A functional behavior assessment (FBA; Dufrene & Lundy, 2019) is typically conducted, and FBA data are used to develop a behavioral intervention plan that consists of antecedent and consequent strategies that reduce the probability of problem behaviors and increase the probability of appropriate replacement behaviors. Additionally, progress monitoring and feedback to students are more frequent than in Tiers 1 and 2. Overall, these levels of supports aim to increase both class-wide and individual student appropriate behaviors while simultaneously decreasing disruptive behaviors in the classroom and have been extensively studied with beneficial results. However, it may be

particularly important to identify effective class or small-group strategies to reduce the number of individual students who are referred to Tier 3, thus reducing the intensity and effort required of individual teachers and school systems.

PBIS in Preschool. Researchers have not tested PBIS in preschool as extensively as other school settings (e.g., elementary and high schools). However, it has been suggested that these strategies may also be effective in early childhood education and preschool settings with only minor adjustments (e.g., age-appropriate language for behavioral expectations; Stormont et al., 2005). Carter and Pool (2012) agreed that modifying expectations to be developmentally appropriate to preschool-aged children is important for preschool PBIS implementation and further suggested reducing the number of broad expectations (e.g., two to four) implemented program-wide and using lesson plans to teach and model those expectations.

Token economies. One class-wide (Tier 1), or targeted (Tier 2) approach that may be utilized is the implementation of a token economy, which provides rewards for appropriate behavior (Fisher et al., 201). Token economies have been studied for decades and have been generally shown to be effective (Doll et al., 2013). Although there have been a number of variations of the token economy, the key feature is the immediate delivery of a tangible, 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 by the individual for a backup reinforcer, typically from a reward menu of items pre-determined for their potential reinforcing effects for the individual. The key benefit to 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 a reinforcer (Doll et al., 2013; Fisher et al., 2011). Another benefit to the token economy is its utility in both the behavior management of an individual client or a group of individuals (e.g., class wide; Drabman et al., 1974; Filcheck et al., 2004; Klimas, 2007; McGoey & DuPaul, 2000; Reitman et al., 2004). Thus, token economies have been applied to a variety of settings (e.g., institutions, jobsites, and classrooms) and populations (e.g., typically developing, developmentally delayed, children, adults).

Reitman and colleagues (2004) utilized an alternating treatment design with a reversal to evaluate and compare the effects of an individual- and group-based class wide token economy system within a Head Start preschool classroom. Within this classroom, three individual students were chosen as target students based on meeting criteria for behavioral referral (i.e., teacher and behavioral screener referrals). Across both types of treatment conditions, a visual token chart system was utilized. The system consisted of a visual representation of seven behavioral levels; top levels indicated good to excellent behavior, middle levels indicated acceptable levels of behavior, and lower levels indicated poor behavior. Levels were moved up based on observations of appropriate behavior. This token economy system also utilized a response cost procedure in which levels were moved down based on observations of inappropriate behavior. Performance at the top levels (i.e., good to excellent behavior) by the end of the session provided students with an opportunity to throw a Velcro ball at a rewards chart; the reward the ball touched or attached to was the earned reward for that session. During individual-based sessions, the opportunity to earn a reward was based on a target student's behavior, and

during group-based sessions, that opportunity was based on randomly selected other students. Results showed that for two of the three participants, rewards earned based on the behavior of an individual student was more effective at reducing disruptive behavior of the target students compared to the group phase (i.e., rewards earned based on the behavior of randomly chosen students). However, the authors noted several limitations to the study including varying levels of teacher-rated treatment acceptability and low rates of teacher-provided praise. Further, the authors failed to collect data on aggregate class wide levels of behavior, so the extent to which either token economy system affected the overall levels of disruptive in the classroom are unknown.

Filcheck et al. (2004) evaluated the effects of the Level System, another levels-based class-wide economy, on the inappropriate behavior of a preschool classroom with 17 children. The level system utilized in this study was similar to Reitman et al. (2004) in that higher levels of the system resulted in children earning access to pre-determined rewards (e.g., quick activity, stickers) and lower levels were not associated with the ability to earn a reward. Children were provided with their own shape on the levels chart, and each child earned a reward based on his or her own behavior (i.e., individual rather than group-based contingency). The teacher also provided verbal praise to children when they exhibited behavior that warranted an increase in their level. Similar to Reitman et al. (2004), this system also utilized a response cost procedure in which verbal warnings were provided to children that exhibited inappropriate behavior and lowered levels following subsequent exhibition of inappropriate behavior. Results of this study showed that inappropriate behavior of the children was on a decreasing trend throughout the Level System phase, with mean frequencies of inappropriate behavior decreasing from

0.45 to 0.29 per minute for baseline and Level System phases, respectively. Further results showed that the Level System phase procedures increased teacher labeled praise statements from 0.07 to 0.50 per minute for baseline and Level System phases, respectively. However promising, the authors noted several limitations to consider when interpreting these results, including low treatment integrity of the token economy procedures. Further, as stated, overall levels of inappropriate behavior were on a decreasing trend throughout the study, including during the withdrawal phase; thus, it may be possible that the decrease in inappropriate behavior may be due to other factors present in the environment (e.g., maturation).

Although the above literature review outlined several studies that implemented variations of a token economy that resulted in positive effects on student inappropriate or disruptive behavior, there are limitations of the current literature base that warrants further scientific evaluation. First, across both treatment strategies, 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). Especially with the growing emphasis on early intervention strategies (Feil et al., 2016; Fox et al., 2002; Stormont, 2002; Webster-Stratton & Hammond, 1998) 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 (e.g., Filcheck et al., 2004; Reitman et al., 2004), and the effect of other variations within this setting should be further evaluated.

Recently, Maggin et al. (2011) and Soares et al. (2016) conducted meta-analyses and design standards reviews of the token economy in schools literature. Meta-analyses

included calculating effect sizes to quantitatively synthesize the findings of studies and design standards reviews included evaluating the methodological rigor of studies using standards described by the What Works Clearinghouse (WWC; Kratochwill et al., 2010). Maggin et al. (2011) was purportedly the first meta-analysis conducted on token economies in the school literature that evaluated the quality of methodological rigor of the included studies. The study included a total of 24 studies that evaluated the effect 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 levels. However, the evaluations on the quality of the studies showed several weaknesses that do not support token economies as an evidence-based practice, including failure to meet WWC design standards (Kratochwill et al., 2010). Soares et al. (2016) results were similar to Maggin et al. (2011) in that token economies produced overall improvements in student behavior across the 28 included studies. In fact, approximately 68% and 25% of studies produced large and medium effect sizes, respectively. Soares et al. (2016) also evaluated the overall quality of the included studies and results showed the number of studies in this body of literature that demonstrate 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) produced similar overall findings that token economies implemented in school settings do 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 included in these meta-analyses. In fact, Maggin et al. (2011) only included K-12 in the inclusion criteria for their meta-analysis 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 of 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 less design standards.

Purpose of the Current Study

The purpose of the current meta-analysis was to determine the effect size of token economies implemented within the preschool setting in single case design studies. Additionally, this study included an evaluation of 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 student behavior?
2. Is the effectiveness of token economies on preschool student 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?

CHAPTER II -METHOD

Literature Search

A literature review was conducted using a multi-step process, ensuring the included articles for the meta-analysis were most appropriate to the current research questions. First, the author used electronic databases relevant to applied psychology available within the author's current institution: APA PsycInfo and Psychology and Behavioral Sciences Collection. Parameters of the initial literature review included a limitation on publishing year and specific keywords. Within the database search, all studies published after 1980 were included. The rationale to limit the range of years followed the one described by Soares et al. (2016) and only included studies published after the passage of Public Law 142 in 1975 which set forth policies and laws related to free appropriate public education to children with disabilities.

Second, three groups of keywords were searched within the databases using Boolean Operators to target the search to more applicable studies. Within-group terms utilized the Boolean Operator "OR" and between-group terms utilized the Boolean Operator "AND": "preschool" or "early childhood" or "head start" or "prek" or "pre-k" AND "token economy" or "tokens" or "token" or "token system" AND "classroom."

Third, the author applied inclusion and exclusion criteria to the initial literature review. Articles were included for the meta-analysis if they met the following inclusion criteria: 1) the study utilized single-case design, 2) the study participants were preschool-aged (2 to 5 years old), 3) the study was conducted in the preschool setting, 4) the study evaluated the effect of token reinforcement on student behavior, 5) the study was published in a peer-reviewed journal, and 6) the study was available in English. The

references for the articles were searched to identify any additional articles not included in the results of the original database search. The author reviewed each citation and identified potentially relevant articles. Next, the abstracts of those articles were reviewed to determine if the study met the aforementioned inclusion criteria. Finally, relevant articles were reviewed in full to determine the extent to which they met inclusion criteria.

Article Coding

Each article was coded for four general categories, including 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 (WWC, 2020) 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 (WWC, 2020), the phase length and minimum data points per phase is grouped into one standard, the standard was separated into two variables for the purpose of this meta-analysis.

For participant characteristics the following variables were coded: whether or not 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 or not 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 or not 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 to run moderator analyses to determine whether or not specific variables moderate or impact the effectiveness token economies may have on the behavior of preschool students. Moderator variables included: Design type, setting, components, interventionist status, percent 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

In order to calculate effect sizes, software was utilized to extract the numerical data for each included article. DigitizeIt Version 2.5 (Bormann, 2012) was used to extract each data point from an image of the graphs for each article. DigitizeIt has been found to be a reliable and valid software package for extracting data (Rakap et al., 2016). 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, data points that contained a negative value were changed to 0. Negative values were determined to result from errors of clicking slightly below the x axis.

Interrater Agreement

The author trained a secondary reviewer on the steps to perform the literature review for the current meta-analysis. The secondary reviewer was a Master's-level behavioral health nurse educator with experience in conducting systematic literature reviews. Two literature reviews were conducted independently by the primary and secondary reviewers. During the initial database search utilizing the Boolean Operators, searches by both reviewers produced the same number of initial articles ($k = 42$). Inclusion criteria were then applied to the 42 articles independently by the reviewers. Agreement in this stage was 91.67% using total count agreement (primary reviewer $k = 11$; secondary reviewer $k = 12$). The reviewers discussed discrepancies until 100% agreement was reached ($k = 11$).

The author developed a coding scheme and trained a secondary coder on coding of all 32 variables for the current meta-analysis. The secondary coder was a school psychology doctoral student with experience in coding and meta-analyses. Training consisted of the primary and secondary coder reviewing the coding scheme and clarifying any questions the secondary coder had regarding definitions of codes. The two coders then practiced coding on an article excluded from the meta-analysis due to failing to meet all of the inclusion criteria. Using an excluded article ensured enough similarity between the practice article and the final included studies (e.g., similar dependent variable, similar design type, etc.). Discrepancies in practice coding were discussed until 100% agreement was met on the practice article.

The author created label codes for the 10 articles included in the current meta-analysis and used a random list generator available online to identify articles to be sent to the secondary coder. Articles were randomized, and the first 3 were chosen for secondary coding for 30% of the included articles. Coding agreement utilized an extract agreement method across variables. For each variable, the coders had to agree on the specific code; 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 an agreement was made. Eleven variables fell under this criterion and coding was discussed. The primary and secondary coders recoded those 11 variables and exact agreement was recalculated and reached 100% agreement.

The secondary coder also extracted data utilizing the aforementioned data extraction method (i.e., Digitize It) for 30% of the articles. Data extraction agreement consisted of the secondary coder independently extracting the data for 30% articles. For data extraction, agreement 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. Initially, each datum for both the primary and secondary coder were rounded to the nearest tenth. Exact agreement across studies averaged 21.64% (range = 14.29% - 28.67%). The primary and secondary coders discussed agreement and discrepancies and determined that the exact agreement may be too stringent for the current data extraction method (i.e., Digitize It); thus, each datum was then rounded to the nearest whole number and agreement was recalculated and found to be 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

Baseline-corrected Tau (Tarlow, 2017) is an effect size statistic that is appropriate for single case design studies. The effect size calculation incorporates both overlap of data points between phases as well as any present baseline trend. Phase data are entered into an online calculator (Tarlow, 2016) for a two-step process. First, the calculator analyzes the baseline data for trends. Second, if the data indicate a significant trend in the baseline, a correction to account for the trend is applied prior to calculating Baseline-Corrected Tau; if the data do not indicate a trend in the baseline, no correction is needed

and Tau (without baseline corrected) is calculated. 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).

Hedge's g

As a second measure of effect size, Hedge's g was also calculated for each study and across studies to produce an omnibus effect size. Hedge's g is based on Standardized Mean Difference (SMD) which is a common parametric statistical method for calculating effect size that can be used for single-case design studies. SMD and Hedge's g is appropriate for comparing two phases (i.e.., phase contrast) and distributes weight to reduce the influence of unequal observations across the two phases (Durlak, 2009). 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).

Data analysis

For baseline-corrected Tau, a free calculator available online (Tarlow, 2016) was utilized to calculate the effect size. First, data for each phase contrast were pasted into the online calculator. Phase contrasts most relevant to the current meta-analysis were determined by the author; generally, A-B contrasts were utilized where A was a baseline phase and B was a treatment phase (Parker & Brossart, 2006). Of note, maintenance or follow up data were not included in phase contrasts for the current meta-analysis. Next, the calculator automatically evaluated the data to test for any significant trends in the baseline data. 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.

To prepare the raw data for calculating Hedge's g, the author calculated the mean and standard deviation for each phase of each study using Microsoft Excel. The phase contrasts that were utilized for Hedge's g matched the phase contrasts used for baseline-corrected Tau (i.e., baseline or withdrawal phases to adjacent treatment phases). The mean and standard deviation calculations for the phase contrasts of the included studies were then entered into R (Harrer et al., 2019a; R Core Team, 2013), which is a free software package that can be used for statistical and graphical analyses. Within R, the *dmetar* package was utilized (Harrer et al., 2019b). Due to differences in sampling across studies, a random effects model was utilized to calculate the omnibus effect of token economies on preschool student's behavior.

CHAPTER III - 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. The author reviewed each abstract, and articles were excluded if they failed to meet any of the 6 inclusion criteria. The remaining article 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. One study (Wolfe et al., 1983) was excluded from the final analyses once in the coding phase because the graphical representation of the data was presented in a way that precluded data extraction using the current methods (i.e., data were presented as only a line or data path without the ability to differentiate between individual datum). In total, 10 articles were utilized for the current meta-analysis (Conyers et al., 2004; Conyers et al., 2003; Filcheck et al., 2004; McGoey & DuPaul, 2000; Miller et al., 1981; Plavnick et al., 2010; Reitman et al., 2004; Sran & Borrero, 2010; Swiezy et al., 1993; Tiano et al., 2005).

Descriptive Statistics

WWC Design Standards

None of the included studies met WWC Version 4.1 (WWC, 2020) design standards overall. In other words, each study failed to "Meet without Reservations" on at least one design standard variable. However, two studies met all criteria with reservations (Conyers et al., 2004; Reitman et al., 2004). All of the included studies met design standards for data availability and systematic manipulation (Conyers et al., 2004;

Conyers et al., 2003; Filcheck et al., 2004; McGoey & DuPaul, 2000; Miller et al., 1981; Plavnick et al., 2010; Reitman et al., 2004; Sran & Borrero, 2010; Swiezy et al., 1993; Tiano et al., 2005). Only 30% of the studies met the design standard regarding IOA (Conyers et al., 2004; Reitman et al., 2004; Tiano et al., 2005). The design standard related to residual effects was met by 66.67% of studies of which this design standard was applicable (Conyers et al., 2004; Reitman et al., 2004). Eighty percent of the studies met the attempts at intervention effects design standard (Conyers et al., 2004; Conyers et al., 2003; McGoey & DuPaul, 2000; Miller et al., 1981; Reitman et al., 2004; Sran & Borrero, 2010; Swiezy et al., 1993; Tiano et al., 2005). Twenty percent of the studies met the design standards for minimum data points per phase without reservations (Miller et al., 1981; Sran & Borrero, 2010) and 50% of the studies met the design standards for minimum data points per phase with reservations (Conyers et al., 2004; Filcheck et al., 2004; McGoey & DuPaul, 2000; Reitman et al., 2004; Swiezy et al., 1993). See Table 1 for standards met per study.

Table 1 *WWC Design Standards Met Per Study*

	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%

Table 1 (continued)

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.

Participant Characteristics

Across all included studies, data on student behavior were collected across 92 participants; however multiple studies only reported aggregate classwide data rather than individual participants. Most studies (70%) failed to report race or ethnicity of the student participants in each study. Of the three that did report participant ethnicity, all participants were reported to be white or Caucasian for 2 studies (McGoey & DuPaul, 2000; Swiezy et al., 1993), and 88.2% participants were reported to be white or Caucasian for one study (Filcheck et al., 2004). The majority of participants across the included studies showed that 35.26% of student participants were female and 64.74% were male (Conyers et al., 2004; Conyers et al., 2003; Filcheck et al., 2004; McGoey & DuPaul, 2000; Miller et al., 1981; Plavnick et al., 2010; Reitman et al., 2004; Sran & Borrero, 2010; Swiezy et al., 1993); Tiano et al. (2005) did not report gender of student participants. Although all of the included studies took place in a preschool classroom

setting, different types of locations were reported across the set of studies. The majority of studies (60%) took place in a regular, public preschool classroom (Conyers et al., 2004; Conyers et al., 2003; Filcheck et al., 2004; McGoey & DuPaul, 2000; Miller et al., 1981; Sran & Borrero, 2010) while 20%, 10%, and 10% of studies took place in Head Start classrooms (Reitman et al, 2004; Tiano et al., 2005), special education classrooms (Plavnick et al., 2010), and church preschool classrooms (Swiezy et al., 1993), respectively.

Study Characteristics

Of the included studies, 20% utilized a withdrawal design (Filcheck et al., 2004; Tiano et al, 2005), 20% utilized a reversal design (Conyers et al., 2003; McGoey & DuPaul, 2000), 20% utilized a multiple baseline design (Plavnick et al., 2010; Swiezy et al., 1993), and 40% utilized an alternating treatments or multielement design (Conyers et al., 2004; Miller et al., 1981; Reitman et al., 2004; Sran & Borrero, 2010). For the purpose of moderator analyses, withdrawal and reversal designs were included in the same category.

Each study's primary dependent variable was coded into two general categories: inappropriate student behavior or appropriate student behavior. 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. Overall, 60% of the studies used inappropriate student behavior as the primary dependent variable (Conyers et al., 2004; Conyers et al., 2003; Filcheck et al., 2004; McGoey & DuPaul, 2000; Reitman et al., 2004; Tiano et al., 2005) and 40% of the studies used

appropriate behavior (Miller et al., 1981; Plavnick et al., 2010; Sran & Borrero, 2010; Swiezy et al., 1993). See Table 2 for definitions of the specific dependent variables for each study.

Table 2 *Dependent Variable Definitions*

	Behavior Category	Examples
Tiano et al. (2005)*	Inappropriate	Inappropriate behavior (whining, crying, yelling, destructive behavior, aggressive behavior); Noncompliance (failure to comply within 5-s of a teacher command); Off-task (failure to attend to the material or task)
McGoey & DuPaul (2000)	Inappropriate	Inappropriate social behaviors (negative social engagement); Off-task (child looks away from activity or teacher for at least 3 s); disobeying rules (deviation from the rules); tantrumming (yelling, kicking, and/or sulking after a social interaction)
Filcheck et al., 2004**	Inappropriate	Whining, crying, temper tantrums, yelling, destructiveness, negativism, pathological self-stimulation, demanding attention, high-rate behavior, talking out of order, being out of area, or cheating
Plavnick et al., 2010	Appropriate	Appropriate sitting (sitting in a staff-designated location and in a manner instructed by staff with minimal movement for the entire interval); Appropriate vocalizing (talking at or below conversational volume)
Reitman et al., 2004	Inappropriate	Noncompliant, disruptive, negative with the teacher, and negative peer interaction.
Sran & Borrero, 2010	Appropriate	Responses per minute. Responses included tracing numbers and uppercase and lowercase letters with a pencil
Swiezy et al., 1993	Appropriate	Percent compliance. Compliance included initiation or completion of the response appropriate to the delivered command within 5 s of the command
Miller et al., 1981	Appropriate	Appropriate rest-time behavior: Sitting or lying with at least half of one's body on the rug, not touching another boy or his rug, and no vocalizations nor other noise-making

Table 2 (continued)

Conyers et al., 2004	Inappropriate	Any instance of screaming, crying, throwing objects or using them as weapons, and refusing to comply with a teacher's request within 5 s
Conyers et al., 2003	Inappropriate	Screaming, crying, throwing oneself on the floor, hitting, kicking, property destruction, throwing objects or using them as weapons, and refusing, ignoring, or resisting a staff member's request

Note. An asterisk (*) indicates the study utilized definitions from a coding scheme developed by Jacobs et al. (2000). A double asterisk (**) indicates the study utilized definitions from a coding scheme developed by McNeil et al. (1991)

Half of the included studies utilized a response cost procedure either within the components of the token economy or directly comparing token reinforcement alone to response cost alone (Tiano et al., 2005; McGoey & DuPaul, 2000; Filcheck et al., 2004; Conyers et al., 2004; Reitman et al., 2004). Other treatment components were also coded for some studies, but were not used in any analyses (e.g., components of parent-child interaction therapy were evaluated in Filcheck et al., 2004; choice was evaluated in Sran & Borrero, 2010).

The exchange rate of tokens varied across the included studies. Two studies failed to report the exchange rate (Plavnick et al., 2010; Tiano et al., 2005), 5 studies reported students were able to exchange tokens for a reward once daily (Conyers et al., 2004; Conyers et al., 2003; McGoey & DuPaul, 2000; Miller et al., 1981; Swiezy et al., 1993), and 3 studies reported students were able to exchange tokens for a reward multiple times a day (Filcheck et al., 2004; Reitman et al., 2004; Sran & Borrero, 2010).

Treatment integrity data were reported in 5 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., 2004). However, two of those studies failed to report specific outcome. MgGoey et al. (2000) reported social validity was high (5.1 average across both teachers on a 6-point Likert scale) and Reitman et al. (2004) reported only poor to moderate social validity, depending on the specific student participant.

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). Only one study reported collecting generalization data during the study. Swiezy et al. (1993) evaluated the degree to which their treatment effects in the classroom generalized to the school playground.

The status of each study's interventionist (i.e., the person responsible for implementing the procedures of the token economy) was coded into two categories:

teacher/staff of the preschool classroom or experimenters not staffed by the preschool. One study did not report the status of the interventionist (Conyers et al., 2003). Of the remaining 9 studies, 60% of the interventionists were the preschool classroom's teacher or staff (e.g., teacher's aide; Filcheck et al., 2004; McGoey & DuPaul, 2000; Miller et al., 1981; Plavnick et al., 2010; Reitman et al., 2004; Tiano et al., 2005) and 40% of the interventionists were experimenters (Conyers et al., 2004; Sran & Borrero, 2010; Swiezy et al., 1993).

Effect Size Calculations

Baseline-Corrected Tau

A total of 63 phase contrasts across studies were analyzed to calculate Baseline-Corrected Tau effect sizes. Using the online calculator (Tarlow, 2016) 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 3 for Baseline-Corrected Tau effect sizes across phase contrast within each study.

Table 3 *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	0.215	Moderate
	Damon	BL1-RC	0.537	Moderate
		BL2-TE	0.000	Small
	Mitch	BL1-RC	0.566	Moderate
		BL2-TE	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

Table 3 (continued)

	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	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	0.539	Moderate
		BL1-IN1	0.346	Moderate
		BL2-GR2	0.679	Large
		BL2-IN2	0.680	Large
	Xavier	BL1-GR1	0.396	Moderate
		BL1-IN1	0.693	Large
		BL2-GR2	0.658	Large
		BL2-IN2	0.756	Large
	Tom	BL1-GR1	0.587	Moderate
		BL1-IN1	0.702	Large
		BL2-GR2	0.648	Large
		BL2-IN2	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

Table 3 (continued)

	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, GR = Group, IN = Individual, NO = No Choice, SI = Single Choice, VA = Varied Choice.

Hedge's g

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

Table 4 *Effect Size by Study*

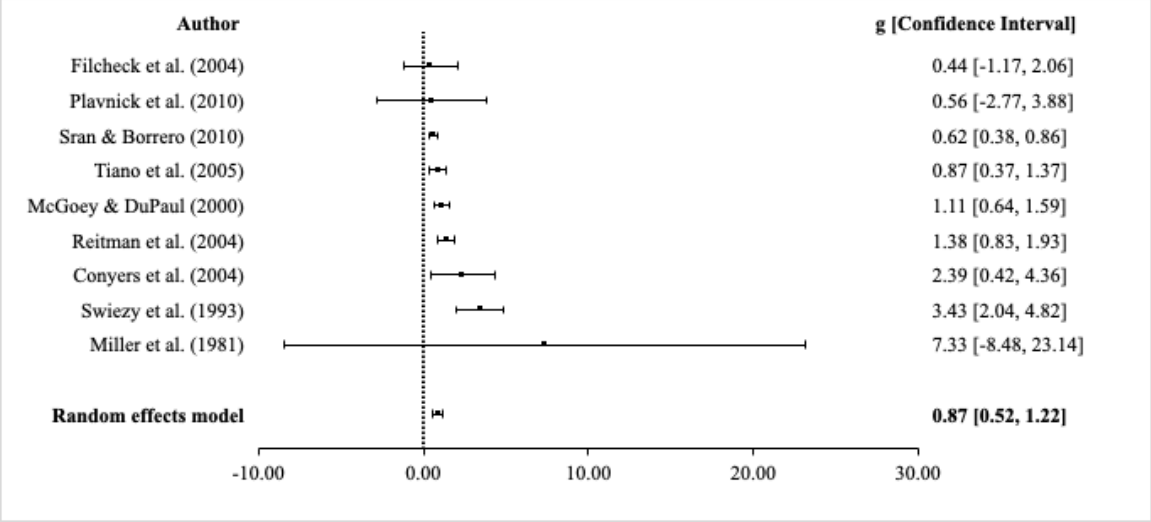
Study	Number of Contrasts	Hedge's g	Confidence Intervals		SE
			Lower	Upper	
Tiano et al. (2005)	6	0.8694^{L}	0.3686	1.3701	0.25548469
McGoey & DuPaul (2000)	16	1.1138^{L}	0.6352	1.5924	0.24418367
Filcheck et al. (2004)	3	0.4425^{S}	-1.1727	2.0576	0.82405612
Plavnick et al. (2010)	2	0.5574^{M}	-2.7681	3.883	1.69670918

Table 4 (continued)

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.

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 of the 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 or not there were outliers present. One outlier was identified (Swiezy et al., 1993) and was removed from analysis 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).

Moderator Analysis

Moderator analyses were conducted for seven variables to determine their effects on the effectiveness 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).

For design type, studies were grouped into three categories: withdrawal/reversal (k = 4), alternating treatments/multielement (k = 2), and multiple baseline (k = 4). Overall, design types produced medium to large effect sizes; however, the effect of design type was not found to be significant ($F_{2, 7} = 3.2236$, $p = 0.1018$).

For setting type, studies were grouped into four categories: Head Start preschool classroom (k = 2), Public preschool classroom (k = 6), Special education preschool classroom (k = 1), and a church-affiliated preschool classroom (k = 1). Medium to large effect sizes were found for each setting. However, the effect of setting was not significant on student behavior outcomes ($F_{3, 6} = 3.7333$, $p= 0.0797$).

For components, studies were categorized as either evaluating token economy and response cost (k = 5) or token economies without the presence of a response cost component (k = 5). Although the presence of a response cost component produced a

larger effect size than token economy alone, the moderator analysis did not find a significant effect on outcome data ($F_{1, 8} = 1.8715$, $p = 0.2085$).

Interventionist status for each study was grouped as either Teacher ($k = 6$) or Experimenter ($k = 3$). Separately, these categories produced large effect sizes. However, the overall effect of interventionist status on student behavior was not found to be significant ($F_{1, 7} = 1.1748$, $p = 0.3143$).

What Works Clearinghouse design standards were used to calculate two different moderator analyses. First, each study was coded overall as either "Met with Reservations" ($k = 2$) or "Does Not Meet." ($k = 8$). It is important to note that no study in the current meta-analysis met full criteria (i.e., "Meets without Reservations") across the separate design standards. The moderator analysis did not produce a significant effect ($F_{1, 8} = 2.1813$, $p= 0.1779$). A separate moderator analysis was conducted with the following groups: Met 66.67% of standards ($k = 3$), Met 83.33% of standards ($k = 5$), met 100% of standards ($k = 2$). Overall, the percentage of design standards was not found to have a significant effect ($F_{2, 7} = 0.9547$, $p= 0.4299$).

The primary dependent variables for each study was categorized into Appropriate Behavior ($k = 4$) and Inappropriate Behavior ($k = 6$). Appropriate behavior produced a medium effect size while inappropriate behavior produced a large effect size. However, the primary dependent variable did not have a significant overall effect ($F_{1, 8} = 1.8735$, $p = 0.2083$). See Table 5 for specific effect size data for each moderator variable.

Table 5

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.

CHAPTER IV - DISCUSSION

The purpose of current meta-analysis was to determine the effect of token economies on student behavior implemented within the preschool setting in single case design studies. Although two recent meta analyses were conducted evaluating the effect of token economies, (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 a favorable and large effect on increasing appropriate student behavior or decreasing inappropriate student behavior in the preschool classroom setting. 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) as inclusion criteria only included k-12 grade levels. 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. However, the number of articles included in the current meta-analysis was 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 5 articles were included in the current meta-analysis and Soares et al. (2016). The inclusion criteria used by Soares and colleagues was limited to the public preschool classroom whereas the current meta-analysis expanded this to other settings (e.g., special education classroom, church-affiliated 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 previous WWC standards (Kratochwill et al., 2010). The current study included a review of design standards with 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, 8 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 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 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 included 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 and results of the moderator analyses should be interpreted with caution given that fewer studies may greatly affect the statistical power necessary to detect differences between groups (Borenstein et al., 2009). Relatively fewer 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.

Limitations

Several limitations of the current meta-analysis should be considered when interpreting the results of the current meta-analysis. First, the initial literature search utilized the two databases relevant to the social and behavioral sciences that were available within the author's current internship institution at the University of Nebraska Medical Center. It may be the case that expanding the search to other databases would have yielded a higher number of articles. However, the ancestral search was used to include articles not otherwise available in the initial search. 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 the case 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). Third, the author coded the dependent variables into two general categories (appropriate and inappropriate student behavior). However, as seen in Table 2, the specific definitions 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). Similarly, 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 including a number of different types of outcome variables may produce different effects. 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. Lack of these data limits the extent to which future researchers can attempt to replicate these studies and limits the degree to which the studies' findings can translate from sample to population. Further, many studies did not report sufficient data related to treatment integrity and social validity. Finally, maintenance and generalization data were not collected for the majority of studies; thus, it is unknown if treatment effects maintain over time and generalize to other settings.

Future Directions

While the current meta-analysis produced results in favor of the overall effectiveness that token economies have on student behavior in the preschool classroom, future studies should tend to aforementioned limitations. 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

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. doi: 10.1901/jaba.1969.2-119

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

Bormann, (2012) DigitizeIt (version 2.0) Retrieved from 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.

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. doi: 10.1111/j.1469-7610.1990.tb00831.x

Carr, E. G., Dulap, G., Horner, R. G., Koegel, R. L., Turnbull, A. P., Sailor, W., … Fox, L. (2002). Positive behavior support: Evolution of an applied science. *Journal of Positive Behavior Interventions, 4,* 4-16. doi: 10.1177/109830070200400102

Carter, D. R. & Pool, J. L. (2012). Appropriate social behavior: Teaching expectations to young children. *Early Childhood Education Journal, 40*(5), 315-321. doi: 10.1007/s10643-012-0516-y. doi: 10.1007/s10643-012-0516-y

Chafouleas, C. M., Riley-Tillman, T. C., & Sassu, K. A. (2006). Acceptability and reported use of daily behavior report cards among teachers. *Journal of Positive Behavior Interventions, 8*(3), 174-182. doi: 10.1177/10983007060080030601.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101. doi: 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. doi: 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. doi: 10.1300/J019v25n03_01

Dufrene, B. A. & Lundy, M. P. (2019). Functional behavior assessment. In K. C. Radley & E. H. Dart (Eds.), *Handbook of behavioral interventions in schools: Multi-tiered systems of support* (pp. 89-105). New York, NY: Oxford University Press. doi: 10.1093/med-psych/9780190843229.003.0006

Durlak, J. (2009) How to Select, Calculate, and Interpret Effect Sizes. Journal of Pediatric Psychology. March: 34(9):917-28.

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. doi: 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. doi: 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. doi: 10.1177/10634266020100030301

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019a). Doing Meta-Analysis in R: A Hands-on Guide. Retrieved from 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. Retrieved from http://dmetar.protectlab.org/.

Horner, R. H., Sugai, G., & Anderson, C. M. (2010). Examining the evidence base for school-wide positive behavior support. *Focus on Exceptional Children, 42*(8), 1-14. doi: 10.17161/foec.v42i8.6906

Horner, R. H., Sugai, G., & Fixsen, D. L. (2017). Implementing effective educational practices at scales of social importance. *Clinical Child and Family Psychology Review, 20*, 25-35. doi: 10.1007/s10567-017-0224-7

Jacobs, J. R., Boggs, S. R., Eyberg, S. M., Edwards, P. D., Querido, J. G., McNeil, C. B., & Funderburk, B. W. (2000). Psychometric properties and reference point data for

the revised edition of the School Observation Coding System. *Behavior Therapy, 31*, 695-712.

Jackson, D. & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods, 8*, 290-302. doi: 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. doi: 10.1037/0033-2909.102.2.187

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.*

LaBrot, Z. C., Dufrene, B., Radley, K. C., & Pasqua, J. (2016). Evaluation of a modified Check-in/Check-out intervention for young children. *Perspectives, 1*, 143-165.

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. doi: 10.1016/j.jsp.2011.05.001

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. doi: 10.1037/h0088790

McIntosh, K. & Goodman, S. (2016). *Integrated multitiered systems of support: Blending RTI and PBIS.* New York, NY: Guilford Press.

McNeil, C.B., Eyberg, S., Eisenstadt, T.H., Newcomb, K., & Funderburk, B. (1991). Parent-Child Interaction Therapy with behavior problem children: Generalization of treatment effects to the school setting. *Journal of Clinical Child Psychology, 20*, 140–151.

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. doi: 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. doi: 10.1177/001440290407100104

Office of Special Education Programs, Technical Assistance Center on Positive Behavioral Interventions and Supports. (2015). *Positive behavioral interventions and supports (PBIS)implementation blueprint: Part 1. Foundations and Supporting Information*. Eugene, OR: University of Oregon.

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

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. doi: 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. doi: 10.1097/00001163-200601000-00004

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from 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. doi: 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. doi: 10.1300/J019v26n02_02

Sabol, T. J. & Pianta, R. C. (2012). Recent trends in research on teacher-child relationships. *Attachment & Human Development, 14*(3), 213-231. doi: 10.1080/14616734.2012.672262

Silver, R. B., Measelle, J., Essex, M., & Armstrong, J. M. (2005). Trajectories of externalizing behavior problems in the classroom: Contributions of child characteristics, family characteristics, and the teacher-child relationship during the school transition. *Journal of School Psychology, 43,* 39-60. doi: 10.1016/j.jsp.2004.11.003

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. doi: 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. doi: 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

Stormont, M., Lewis, T. J., & Beckner, R. (2005). Positive behavior support systems: Applying key features in preschool settings. *Teaching Exceptional Children, 37*(6), 42-49. doi: 10.1177/004005990503700605

Sugai, G. & Horner, R. H. (2014). Positive behavior support, school-wide. In C. R. Reynolds, K. J., Vannest, & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education* (pp. 7-10). Hoboken, NJ: Wiley. doi: 10.1002/9781118660584.ese1902

Sugai, G. & Horner, R. R. (2006). A promising approach for expanding and sustaining school-wide positive behavior support. *School Psychology Review, 35*(2), 245-259.

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

Tarlow, K. R. (2016). Baseline corrected tau calculator. Retrieved from 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. doi: 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. doi: 10.1037/h0100298

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

Tingstrom, D. H., Sterling-Turner, H., & Wilczynski, S. (2006). The Good Behavior Game: 1969-2002. *Behavior Modification, 30*(2), 225-253. doi: 10.1177/0145445503261165

U.S. Department of Education. (2020). *Forty-second annual report to Congress on the implementation of the Individuals with Disabilities Education Act*. Washington, DC: 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. doi: 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. doi: 10.1002/jcad.12038

Webster-Stratton, C. & Hammond, M. (1998). Conduct problems and level of social competence in Head Start children: Prevalence, pervasiveness, and associated risk factors. *Clinical Child and Family Psychology Review, 1*(2), 101-124. doi: 10.1023/A:1021835728803

What Works Clearinghouse. (2020). What Works Clearinghouse Standards Handbook, Version 4.1. Washington, DC: 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