Reinforcement Practicality for Middle School Students: A Meta-Analysis

Kelly C. Dreger, Ed.D.

Steve Downey, Ph.D.

Valdosta State University

The need for evolving support interventions that can help students in a wide range of settings is an ongoing requirement for middle schools today. Token reinforcement, which is a form of extrinsic motivation and incentivization, is studied within this meta-analysis to determine if significant treatment effects exist overall and if there are studies that show more gains than others. Most studies report significant positive gains individually, but the statistical significance is lost when the studies are reviewed as a whole. Variables such as sample size requirements, treatment effect variation, and session time all influence treatment effect size. Reinforcement has been shown to be a viable strategy for differentiation, but the area of standardization has yet to be adequately addressed within past and present research. Some effect size traits reported from the literature are supported within this meta-analysis, but the sampling, analysis, and interpretation protocols exhibited by certain studies make it difficult to remove bias and confounding within reinforcement studies. Further research avenues and additional considerations are discussed.

Keywords: Incentives, Reinforcement, Instruction, Standardization, Differentiation

Introduction

K-12 education has its set of issues that require more practical solutions. Teaching jobs are typically underfunded because of significant increases in salary and public education cuts (Weingarten, 2019). Teacher optimism has decreased over time, given the problematic state of affairs (Houghton Mifflin Harcourt, 2019; Weingarten, 2019). Teachers are looking for more sustainable systems in addition to what they already use (Abramovich, Schunn, and Higashi, 2013; DeFrancis, 2016; McClintic-Gilbert, Corpus, Wormington, and Haimovitz, 2013; Tan, Kasiveloo, & Abdullah, 2022). For instance, nearly 20% of schools that have school-wide information systems utilize suspensions as a main consequence of behavior, resulting in over 10,000 suspensions overall (Eliason, Horner, & May, 2013). There is demand for more innovative social, emotional, academic, behavioral, and technological supports that would help students succeed in the classroom (Bureau et al., 2022; National Technical Assistance Center on Positive Behavior Interventions and Support, 2017; Shakespeare, Peterkin, & Bourne, 2018; Simonsen et al., 2008). There is also the need for the adaptation and improvement of more traditional methods to address support frameworks such as Response to Intervention (RTI), Positive Behavior Interventions and Supports (PBIS), Multi-Tiered System of Support (MTSS), and other practical, relevant decision-making frameworks. Adaptation into a school-wide system has to account for such budgetary concerns as number of schools, training, personnel allocation, data collection, tier levels affected, and competition with alternative initiatives (Swain-Bradway, Lindstrom Johnson, Bradshaw, & McIntosh, 2017; Horner et al., 2012).

One possible intervention system that can be improved and differentiated as stipulated is the reinforcement system. In this article, the practical applications of token reinforcement as a strategy will be discussed according to results from a meta-analysis on these specific incentives. This meta-analysis summarizes practical incentives, specifically as it applies to teaching and learning. It helps provide standardization techniques for motivation within research and practice. It also provides a statistical basis for support of findings within previous literature.

Research questions were established in order to determine further information about studies reviewed within Dreger (2017). Four questions were established for this particular investigation: (A) What is the magnitude of the overall effect size for participation in reinforcement interventions within middle schools? (B) To what extent, if any, does variation exist within the effect sizes given for reinforcement programs in middle schools? (C) To what extent is practical significance influenced by intended program outcomes (i.e., performance, behavior, and motivation)? (D) Which study interventions for middle school students showed the most gains based on the effect sizes given? For the purposes of this study, token reinforcement refers to an object or symbol that is exchanged for goods or services (Hackenberg, 2009). This can include such items as points, money, tallies, grades, and cards. Tokens have been and continue to be used in traditional, hybrid, and online educational settings.

Conceptual Framework

The strategy of reinforcement has been around for some time. Historical applications of token and other tangible reinforcers, in particular, have been around since 8000 BC (Schmandt-Besserat, 1992). Official psychological terminology dates back as early as the 1930s, and educational management of them goes as far back as the 1960s (Doll, McLaughlin, & Barretto, 2013; Gaughan, 1985; Hackenberg, 2009; Taylor, 2000). Despite the rich history available, there are problems that still exist in its application (Tan et al., 2022; Branch, Reid, & Plutzer, 2021). The very nature of the topic requires an examination of students' needs and environmental circumstances by teachers and researchers alike; therefore, it is important to get to what actually works and discover the degree to which reinforcement can be used to optimize learning, if such a degree exists. There is a gap that still exists between what makes for viable theory versus what makes for good practice in schools. The studies themselves have different frameworks, including those within the areas of operant learning, behavioral analysis, social learning, and achievement goal development. What they all have in common is the following: (A) A motivational outcome needs to be reached; (B) Events are planned in relation to the outcome; (C) Students behave according to what happens within the planned events; (D) Participants receive a particular consequence for their behavior; and (E) Plans are either modified or maintained as a result of the consequence received.

Not all teachers agree that incentives are useful for students, and a meta-analysis presents the overall results of rigorous, evidence-based, and data-driven studies about reinforcers. The term meta-analysis was initially created in 1976 by Gene Glass (Rickert, 2014). A meta-analysis is a systematic approach of statistical analysis that can be viewed in two essential ways: a) a comprehensive review and statistical analysis of studies that goes beyond a typical research paper or literature review or b) a research method that is used to integrate the different perspectives and findings of studies (Glass, 1976; Denson & Seltzer, 2011; Newman, 2003). Meta-analyses are implemented to answer questions about magnitude, precision, variation, and compatible effects (Denson & Seltzer, 2011; Hicks, 2014; Higgins, 2008). They combine data across a variety of studies, which may include experimental data, survey data, or other relevant statistical information (Denson & Seltzer, 2011; Tojo, 2013). Regardless of how one views meta-analyses, the framework for completion is as follows: (A) Develop research questions based on a topic of interest; (B) Conduct a search on the relevant topic and include essential search terms; (C) Find important resources about the topic to determine a final set; (D) Retrieve essential information from the chosen resources; (E) Determine the quality of the resources; (F) Summarize the heterogeneity of the resources; (G) Determine effect size, appropriate models, and forest plots; (H) Determine if publication bias exists and create a funnel plot; and (I)

Conduct analyses, including subgroup analyses and regression, that sufficiently answer the research questions. Studies were coded and categorized according to basic requirements for procedures recommended by Basu (2017), Denson and Seltzer (2011), Del Re (2015), Maksimović (2011), and Ahn, Ames, and Myers (2012). All coding methods involved determining a) general study characteristics, b) essential definitions related to the problem, c) the search process, d) study quality, and e) the reporting of results. All of this information was used to determine which studies would be used for this meta. Basu (2017) provided a step-by-step outline of conducting a basic meta-analysis. Del Re (2015) did this as well, but more emphasis was placed on the statistics and programming aspects of performing one. Recommendations given by Denson and Seltzer (2011) and Maksimović (2011) were based on what was acceptable for meta-analyses and hierarchical modeling within education. Ahn, Ames, and Myers (2012) based their suggestions on a plethora of validity frameworks, including Cooper's (1982) checklist for validity of meta-analysis, Valentine, Cooper, Patall, Tyson, and Robinson's (2010) application of Cooper's checklist, the meta-analysis reporting standards developed by the APA (2008), and the MUTOS framework initially proposed by Cronbach (1982) and modified by Aloe and Becker (2009).

Methodology

To start the meta-analytic process, search terms were perused in order to confirm what was relevant to the current meta-analysis. The search terms were based on an electronic concept map that was created for this investigation, which elaborated on previous hand-written and electronic concept maps created by the authors concerning token reinforcement. Figure 1 shows an electronic concept map for this study.

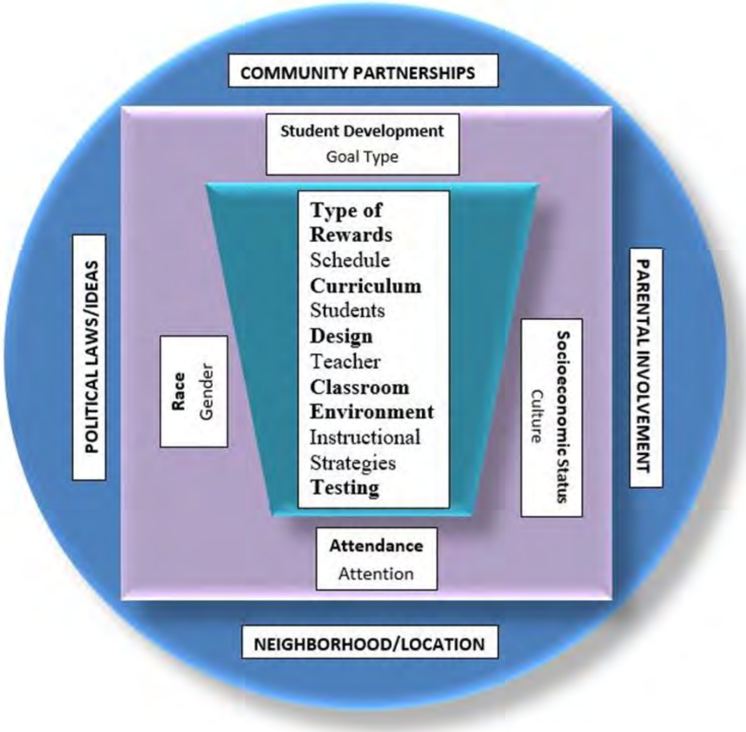

Figure 1. Electronic concept map for token reinforcement investigation.

After the study pool was determined, information had to be retrieved in order to find what was feasible for a meta-analysis. The studies were sorted according to methodological structure similarities. This provided a logical method of organization due to the fact it took into account the actual content of the various resources. The classifications found were the following: (A) Surveys and Questionnaires, (B) Meta Discussions, (C) Experiments and Quasi-experiments with Students, (D) Experiments with (Non-Human) Animals, (E) Code of Conduct and Ethics Manuals, (F) General Papers and Reports, (G) Strictly Qualitative Case Studies and Interviews, (H)Literature Reviews, (I) Books, and (J) Correlational and Ambiguous Effect Studies. Several resources were consulted to determine what actually could fit within the meta, including the ethics handbook from the American Psychological Association (2010), definitions offered by The Cochrane Collaboration (2005), the token methodology suggestions offered by Maggin, Chafouleas, Goddard, & Johnson (2011) as well as the Preferred Reporting Items of Systematic

Reviews and Meta-Analyses (PRISMA) checklist Moher et al. (2009). A synthesis of essential criteria is displayed in Figure 2. The process described in Figure 2 is an overview of the questions asked for this inquiry. The most important question here has to do with the questions and the scope of the project. From there, statistical data-gathering, study integrity, rigor, variable determination, system use, and study participant details were documented in an audit trail journal.

The following characteristics were logically synthesized and analyzed: (A) id number, (B) study type, (C) location, (D) year, (E) outcome data, (F) sample size, (G) effect size, (H) variance, (I) length of sessions, (J) primary grouping variable, and (K) effect size direction. These specific characteristics were placed in tabular format using Excel and R statistical software. The id number is a categorical number assigned to a study that sets it apart from all other studies. Out of the 129 resources found about token reinforcers, there were 31 that had the required information needed for a meta-analysis. Studies were labeled consecutively from 1 to 31. Study type, for the 31 studies, contained the following: a) Experiments, Quasi-experiments, and Causal Comparative Studies with Students, b) Correlational and Ambiguous Methods, and c) Surveys and Questionnaires. Location indicated general geographical groupings of studies: Northeast, Midwest, South, West, Other/Extenuating Circumstances. Year has a few categories: 70s to 80s, 90s to 00s, and 2010s. Outcome data in terms of numbers (i.e., raw scores, means, and standard deviations) were originally sorted within two groups. One group was labeled the control (pretest) group and the second group was labeled the experimental (posttest) group. Outcome data were then categorically sorted into Performance, Behavior, Motivation, or a Combination of measures. Sample size indicated the number of participants within a particular study.

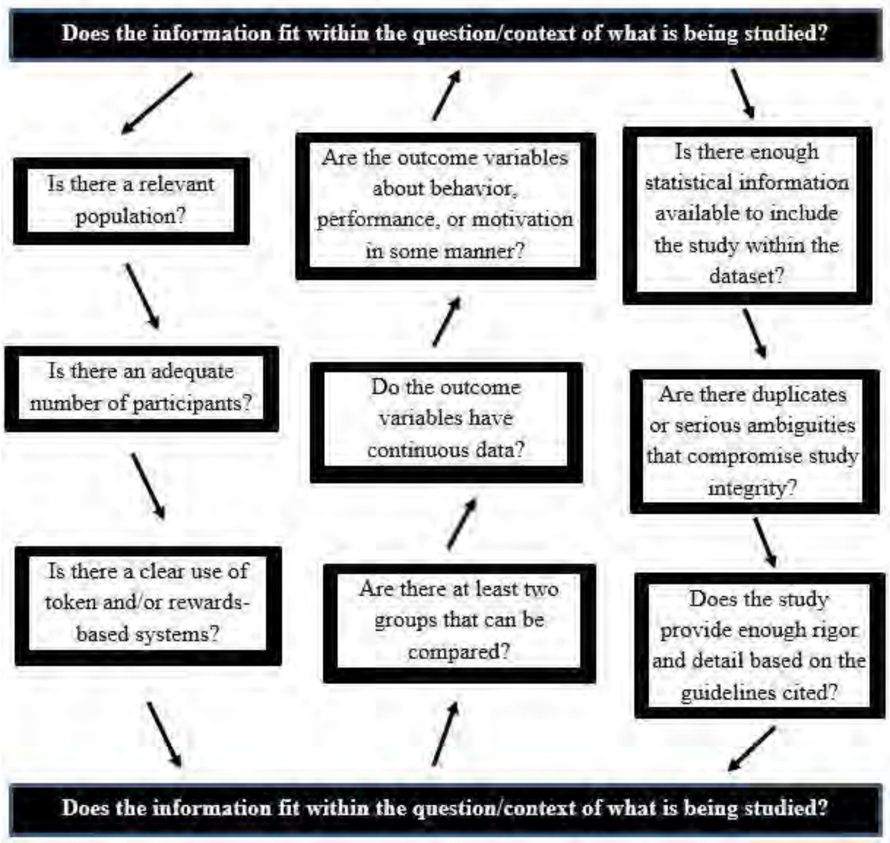
Does the information fit within the question/context of what is being studied?
Is there a relevant population?
Is there an adequate number of participants?
Is there a clear use of token and/or rewards-based systems?
Are there at least two groups that can be compared?
Do the outcome variables have continuous data?
Are the outcome variables about behavior, performance, or motivation in some manner?
Is there enough statistical information available to include the study within the dataset?
Are there duplicates or serious ambiguities that compromise study integrity?
Does the study provide enough rigor and detail based on the guidelines cited?
Does the information fit within the question/context of what is being studied?

Figure 2. Essential criteria used for review and meta-analysis studies pertaining to reinforcement.

Effect size and effect size variance were calculated using formulas for Cohen’s *d* that were converted into Hedges’ *g* for more precise and uniform measurements. The result for Hedges*’ g* was recommended by Del Re (2015) and used within the majority of R calculations. Like the outcome data, both numeric values and categories. Effect size can be sorted as Small, Medium, and Large, which indicates the importance of whatever effect is being objectively measured (Field, Miles, & Field, 2012). This was used along with a Trivial category to indicate something that was non-significant but was still a result. A small effect size would start at .2 when it is shown (Cohen, 1988). A medium effect size would be at .5, and a large effect size would be at .8 in the results. Olejnik (1984) indicates that a small effect size would have a 1%

explained variance. Medium effect sizes would start at 6%, and large effect sizes would have at least 13% explained variance.

The length of the sessions were defined as the following categories: Two Times At Most, Multiple Times, Weeks to Months, and At Least A Year. The primary grouping variable had two categories that labeled if studies were mainly time-based or treatment-based. Most studies were grouped mainly by time (58.06%), and others were grouped mainly by treatment (41.94%). Effect size direction could be positive, negative, or zero (neutral). Direction did not overlap into the Trivial, Small, Medium, and Large categories used for effect size during rank-based modeling. For example, a study with a Large effect size could have an .8 or -.8 as a value. Traits that did not have information available for all studies were discarded. Because the studies did vary in terms of methodology, there were z scores computed for the means used within the outcome data. The outcome data did not show much variation in terms of z-scores; therefore, more complex procedures were employed within R statistical software to determine if any heterogeneity existed for the variables.

Data Analysis

The start date for the documented procedures was July 2018, and the end date was August 2020. For a team of researchers, systematic research and review methods are typically completed within 9 to 24 months (University of Edinburgh, 2013; University of York, 2009). Steps A-E have already been discussed in the previous sections. Steps F-I were completed in R statistical software.

There were 3 Rounds of analyses to complete Steps F-I. The idea of having multiple rounds of data allows for systematic triangulation of all resources in order to provide findings that are credible (Altman, 1991; Creswell, 2009; Maxwell, 2012). Round 1 required the analysis of raw data according to essential assumptions of linear modeling and parametric testing. The assumptions checked involved linearity, collinearity, independence, influential data, normality, and heterogeneity. This gave the data needed to complete Step F. Round 2 involved the analysis of transformed variables to see how all data could be analyzed and fitted properly. Each variable was transformed so that no correlations would be found if the variables were put into a model. Model appropriateness was determined by whether or not all variables could meet assumptions as a whole and in parts. A mixed effects model was determined to be the most appropriate option for the data. The information in Round 2 was crucial to complete Steps G and H. Round 3 involved procedures that clarified previous operations in Rounds 1 and 2. It also contained additional procedures in order to adequately answer the research questions, given the parameters set after Round 2. For instance, correlations were redone in Spearman rho and Kendall tau because Pearson correlations would no longer apply to the variables. Procedures for non-linear, non-parametric modeling that were not addressed in the second round were addressed in this round to conclude analysis. The information in Round 3 completed requirements within Step I.

For Question 1, effect size calculations first required Cohen's *d* to be calculated and converted into Hedges' *g*. Without the conversion, a possible sample bias would have been present in the results. Using the MAd, metafor, and compute.es packages in R yielded the overall effect size and variance needed for Question 1. The method of estimation was Borenstein, Hedges, Higgins, and Rothstein (BHHR). This method helps to combine effect sizes together within each study to create an unbiased effect size estimate for what needs to be aggregated (Del Re, 2015). For Question 2, Effect sizes were generated by suggestions by Del Re (2015), Basu (2017), and Denson and Seltzer (2011) that specifically apply to determining effect sizes for each study, even when there are multiple means given. Question 3 required that the aggregated effect sizes according to outcome were coded where Performance = 0, Behavior = 1, and 2 = Motivation. Then, bivariate correlation testing was conducted, where y was effect size and x was an independent variable. This was able to test influences of effect size. A heterogeneity test was performed to determine variation within the model. Questions 3 and 4 both required moderator and correlation testing to determine if influences and publication biases existed. Question 4 required comparisons of effect sizes to determine which studies had the highest effect sizes. Forest plots and funnel plots were produced in order to identify magnitude sizes and any possible significant biases in the results.

Results

To answer Question 1, the overall effect size for the studies needed to be found. Based on Cohen's *d*, the overall effect size is approximately 1.16, with an effect size variance of 0.08. Using Hedges' *g* shows the overall effect size to be approximately 1.13, with an effect size variance of 0.07. The overall effect size was very large, but there was little variance in treatment effect between the treatment group mean for all studies (*M* = 79.77; *SD* = 116.83) and the control group mean for all studies (*M* = 90.05; *SD* = 133.93). To find the actual significance of the effect size given and check for statistical accuracy, effect size would need to be determined for each of the 31 studies. It is important to note that the magnitude for each study is not the same as the magnitude given overall. Effect sizes for each study are given for Question 2.

To answer Question 2, there needed to be 31 aggregated effect sizes to determine any variation. Table 1 lists means, standard deviations, effect sizes, and variances per study. They are sorted from lowest to highest effect size (*g*) in terms of magnitude, without order in terms of direction.

Table 1

Effect size results for meta-analysis, aggregated per study

Study	CM	CSD	EM	ESD	g	Vg
McClintic-Gilbert et al. (2013)	3.17	0.32	3.17	0.06	-0.018	0.017
Mucherah and Yoder (2008)	2.88	0.18	2.87	0.22	-0.037	0.005
Cross (1981)	76.10	104.01	75.00	105.71	-0.062	0.015
Abramovich et al. (2013)	3.80	0.27	3.78	0.22	-0.080	0.039
Self-Brown and Mathews (2003)	5.49	0.14	6.38	5.68	0.084	0.058
Truchlicka et al. (1998)	81.04	7.52	81.28	7.18	0.101	0.447
Urdan and Midgley (2003)	7.87	0.28	7.90	0.26	0.112	0.004
Wulfert et al. (2002)	25.39	28.08	29.39	35.61	-0.169	0.046
Strahan and Layell (2006)	206.90	60.55	208.38	61.85	0.175	0.019
Ames and Archer (1988)	20.35	28.51	18.35	22.83	-0.215	0.007
Miller (1981)	105.70	2.61	109.59	1.50	0.251	0.030
Hayenga and Corpus (2010)	2.98	0.01	2.84	0.02	-0.279	0.004
Simon et al. (1982)	0.77	0.14	0.81	0.34	-0.329	0.280
Habaibeh-Sayegh (2014)	145.19	185.00	151.09	190.88	0.334	0.024
Gaughan (1985)	2.57	1.79	3.30	1.98	0.383	0.050
Young-Welch (2008)	149.00	10.61	120.50	6.36	0.570	0.066
Borrero et al. (2010)	0.71	0.00	1.45	0.37	0.628	0.289
Hoeltzel (1973)	41.81	46.07	45.51	47.78	0.749	0.266
Taylor (2000)	57.44	61.27	56.63	60.15	0.788	0.047
Marinak and Gambrell (2008)	435.84	60.97	230.98	190.96	-0.806	0.076
Dreger (2017)	79.76	2.07	65.94	3.71	-0.843	0.017

Devers and Bradley-Johnson (1994)	94.03	5.30	103.10	6.87	0.856	0.105
Swain and McLaughlin (1998)	54.50	22.52	83.00	2.16	1.549	0.528
Lynch et al. (2009)	73.67	0.00	91.75	1.90	1.671	0.254
Hansen and Lignugaris/Kraft (2005)	0.12	0.14	0.31	0.03	1.772	0.289
Popkin and Skinner (2003)	46.76	31.13	54.12	37.66	1.796	0.306
McDonald et al. (2014)	23.80	4.50	11.93	5.46	-1.897	0.727
Novak and Hammond (1983)	2.34	0.00	7.33	1.44	3.619	0.549
Unrau and Schlackman (2006)	2.82	0.01	2.71	0.01	-7.419	0.034
Baker and Wigfield (1999)	2.85	0.26	25.94	28.61	11.491	0.166
Yager (2008)	0.16	0.27	0.50	0.16	18.505	1.038

Note. Means, standard deviations, effect sizes, and effect size variances for the 31 studies in the meta-analysis. CM = Control Group Mean; CSD = Control Group Standard Deviation; EM = Experimental Mean; ESD = Experimental Standard Deviation; g = Hedges' *g* effect size; Vg = variance for *g*.

In order to get a better assessment of heterogeneity as a whole, a test was conducted using the mareg function within the Mad package for *R*. From the heterogeneity test, the effect size *g* was approximately 1.04, but there were no significant treatment effects among the means calculated (p = 0.156). The heterogeneity estimator (*QE*) was equal to 2931.771, with statistical significance (p = 0.000). Although no significant treatment effects were found, there was significance in terms of heterogeneity. The confint function within *R* generated a heterogeneity estimate of 99.85%. With transformations, the effect size estimate became 0.156, with significant differences found overall (p = 0.00). The QE moved down to 688.590, with statistical significance (p = 0.000). The estimated percentage of heterogeneity decreased slightly to

99.45%. With the means from the 31 entries, there was little difference overall between the control group (*M* = 56.64, *SD* = 88.08) and the experimental group (*M* = 51.80, *SD* = 62.16). The z-scores calculated helped to support the assertion that means for both groups were not statistically different in terms of treatment effects since the majority of scores had a mean of 0 and a standard deviation of 1. The spread of the scores was very high for the groups, which was not supported by the z-score conversion.

For Question 3, effect sizes were determined from the information on outcome type. Behavior had an approximated effect size where g = 2.38 and a variance where Vg = 0.10. This effect size and variance was the highest of the three outcome categories. There was also a large effect size seen for the Performance outcome category, where g is approximated to be 0.91 and Vg is approximated to be 0.08. A small effect size was seen with aggregated Motivation outcomes, where approximations showed that g = -0.39 and Vg = 0.04. Heterogeneity testing for the outcomes determined the extent of significance for the effect sizes. The test indicated an overall effect size estimated at 0.95, where *p* = 0.24. Although varying effect sizes existed, none of them produced a statistically significant effect where the groups were concerned. Where the significant influence existed was within the QE estimator, which was 58.61 with a p-value of 0. There was an existence of 96.54% heterogeneity, which warranted further investigation since the outcomes did not account for the extremely large differences in the data.

Unlike the heterogeneity test of overall effect size, the moderator testing involved specific testing within each category. Significant *p* values ($p < 0.05$) were found for study type, study outcome, effect size direction, and year of publication. Influential beta weights were found in the effect size ~ outcome equation within the following studies: McDonald et al. (2014), Baker and Wigfield (1999), Yager (2008), and Unrau and Schlackman (2006).

For Question 4, effect sizes were calculated for all studies (See Table 1). There were 31 effect sizes calculated from the information. There were 19 out of 31 studies (61.29%) that showed positive results in favor of the use of token and/or extrinsic interventions. There were 12 studies (38.71%) that had a negative effect size. The study that had the largest effect size gain was Yager (2008). The effect size forest plot, however, showed this as having a small amount of precision when compared to other studies (See Figure 3). The control group in the metadata (*M* = 0.16, *SD* = 0.27) did score lower than the experimental group (*M* = 0.50, *SD* = 0.16). The second largest effect size originated from Baker and Wigfield (1999), where the treatment group (*M* = 25.94, *SD* = 28.61) outperformed the control group (*M* = 2.85, *SD* = 0.26). Thirdly, Unrau and Schlackman (2006) showed a drastic decline in treatment effects between the control group (*M* = 2.82, *SD* = 0.01) and the experimental group (*M* = 2.71, *SD* = 0.01). The plots indicated that the studies are similar in terms of effect size significance when they should not be (See Figures 3 and 4). Bias existed in the sample data, specifically in how the results were interpreted, reported, and selected within past literature. Results indicated that 38.71% of studies reported large effect sizes that were statistically significant, but 29.03% of effect sizes reported did not have any statistical significance. Only two studies (6.45%) actually had the sample sizes required to say that the large effect size could be generalizable. Most outcome data for token reinforcement measured performance-based results (48.39%). Study implementation would generally take place for weeks or months (38.71%).

Furthermore, there were five moderators found that showed varying results. A moderator existed where studies that had surveys or questionnaires had a significantly higher effect size during treatment and control phases (*p* = .007). Another moderator was found in outcome type, where behavior was significantly higher in effect size (*p* = .021) and performance was significantly lower (*p* = .048). A third moderator was effect size direction, where studies with positive effect size direction showed significant improvements over time (*p* = 0.025). In year, there was a possible influence (*p* = 0.047) found with studies made between the 1990s and 2000s, with a significantly higher effect size than studies done in other time periods. Finally, a fifth important moderator was variance. Studies with a higher effect size variance were more likely to have high effect sizes (*p* < .05).

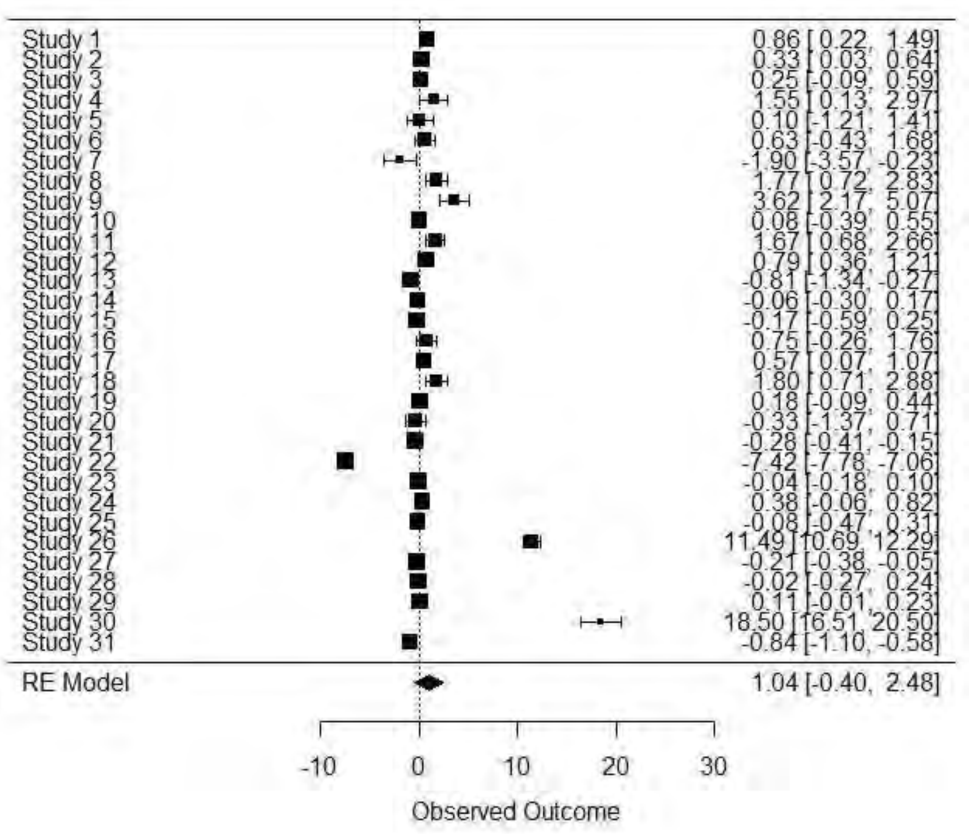

Figure 3. Forest plot for the 31 studies of interest.

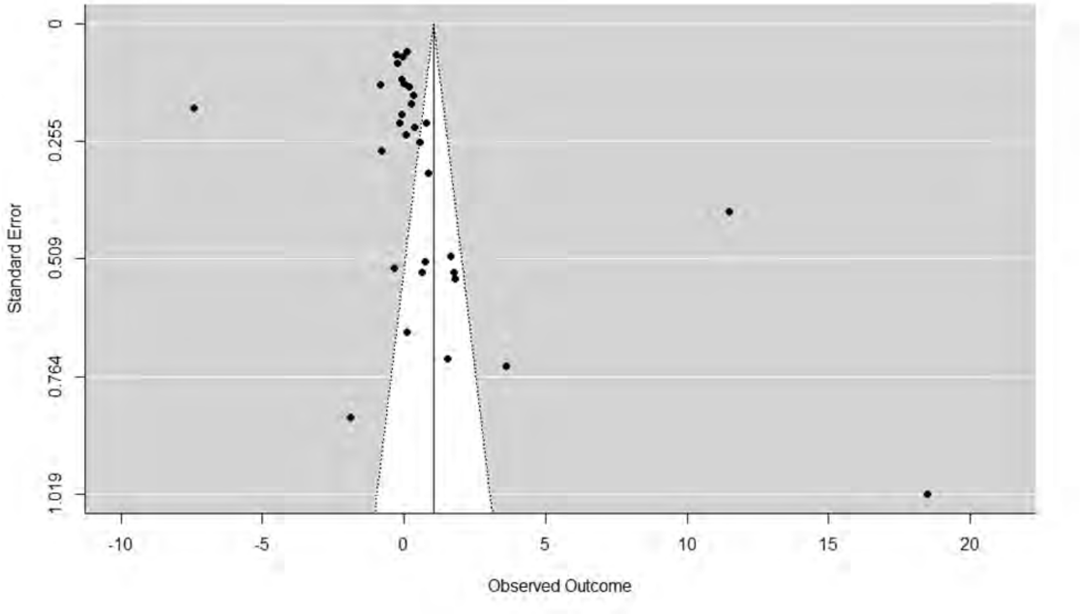

Figure 4. Funnel Plot for Meta-Analysis.

Discussion

What the results for Question 1 tell us is that the treatment effects may be large, but not all variables were significant influences. There was a statistically significant difference in terms of sample size and effect size variation. The outcomes within the data were not statistically significant. This is why z-score standardization showed no significance, but the heterogeneity tests did show significance. For this study, teachers and researchers need to look at how students were grouped in the sample to understand the positive effect sizes. The number of participants can heavily influence the treatment effects. For the Clmm2 models, it is important to remember that large does not mean large sample size, but rather it means the minimum number needed for a statistically large finding. The less precision the results had, the more likely it showed a negative effect size. In other words, studies with low participant numbers and large effect sizes tended to show more unfavorable results. As the sample size coding went up in terms of effect size, the amount of variability in the data tended to increase. This means that less precision indicated more variability, even though minimum requirements in terms of sample size were met.

Results for Question 2 show that the study effect sizes, effect size variances, and design variations are very dissimilar. There were different study designs and requirements, and this is where the variation (i.e., heterogeneity) becomes statistically significant overall. If teachers, researchers, and other stakeholders share the information found herein, they can emphasize that token reinforcement is a strategy that helps with differentiation and instructional flexibility. There are so many ways that reinforcement can be implemented, and it can be used with different types of students. There are clear examples from the literature of how reinforcement can vary according to the needs of students. The problem is pinpointing how it is good in terms of standardization. Sometimes it works and sometimes it does not work for those involved. Conducting a meta-analysis in the way shown in this article is a strategy that helps determine where the strengths are and why. For instance, studies with an effect size that did not exceed 1 in size and direction ($0 \leq x \leq 1$) or go below -1 in size and direction ($-1 \leq x \leq 0$) are more credible from a statistical, quantitative standpoint than those that did.

Question 3 emphasizes that using different tests may yield different results. This is also seen within Question 1. What happens overall in the outcome may not account for what is reported in literature, what happens with individuals, or how variables influence one another within the actual methods. The aggregated data had confounding values (i.e., zero) removed or adjusted according to packages used in R. There were no significant influences from outcome type when the studies were tested as a whole, but bias testing found a different result. In terms of what was seen in the literature, strategies for behavioral outcomes and performance outcomes tended to have significant effect sizes when reported, with behavior strategies having more of a significant impact. Another point made that is also supported by Question 2 is that treatment effect intensity, or magnitude, does not show how good or bad a treatment is. To determine how well the outcomes were when token reinforcement was in use, teachers and researchers have to look at the effect size direction. This is important to note because a teacher or researcher cannot automatically say something is good just because it has a high number attached to it. The intensity of the treatment effect did vary, and the direction varied as well.

When interpreting forest plots in Question 4, it is important to know where the gain or loss would occur. Studies within the forest plot demonstrated that statistical information may not account for all results reported in the literature; however, the numbers do help strengthen the argument for the effects they do account for within the findings. When accurately depicting the results in forest and funnel plots, the untransformed version gives a better visual of the details between and among studies. A possible explanation for the lack of coverage in funnel plots is that studies about reinforcement, including tokens in particular, can be hard to replicate and generalize. The statistical findings, when compared to other studies, may contradict those who strongly favor or oppose the use of reinforcement as a whole.

According to Egger et al. (1997), Banks, Kepes, and Banks (2012) and Sterne et al. (2011), there can be numerous reasons for bias, specifically publication bias, if it exists for sample data. For example, there could be lack of research dissemination, slow dissemination, inadequate decision making, inadequate reporting, withholding of insignificant results, lack of access, and unfair favoritism for more dramatic results. Sedgwick (2014), however, identifies publication bias as “the omission of unpublished trials from a meta-analysis” (p. 1). Using correlation tests and regression tests would determine if any statistically undue influences are found in the data. It is clear from the tests done about collinearity that there are significant relationships within the data to indicate that confounding exists, meaning that not everything was controlled for during the studies' implementation. High variances within the data would indicate a wide spread in scores and more differences found within the data.

To illustrate the similarities and differences between reported results and the results found in the meta-analysis, several noteworthy studies will be discussed: McClintic et al (2013), Yager(2008), Urdan and Midgley (2003), Hayenga and Corpus (2010), Unrau and Schlackman (2006), Baker and Wigfield (1999), and McDonald et al. (2014). The negative effect of extrinsic motivators reported in McClintic-Gilbert et al. (2013) was supported with this meta (*g* = -0.018), but the magnitude of the impact is debatable and actually the least potent when compared to other studies. The positive and intense effects of verbal and tangible incentives seen in Yager (2008) are supported in the meta (*g* = 18.505), but the study itself limits the significantly positive effects to students who tend to need more small group and individualized interventions. It also limits the results to interval scheduling. So even though similar effects are shown in the results, it is not relevant to those who would receive a generalized form of instruction unless the incentive schedule is time-based. If the general education incentive schedule is based on specific behaviors, responses, and outcomes, then the results would not apply. The very large effect size indicated that sample sizes and analysis units were either not accurately reported or were lacking details that would be easily generalizable. This meta did not support the significant results by Urdan and Midgley (2003), which stated that positive changes in mastery goals showed significant improvements in students' outcomes. No significance was found in any of the statistical measures done in the meta-analysis, but the positive effects were supported (*g* = 0.112). Hayenga and Corpus (2010) reported a significantly positive treatment effect with high intrinsic and low extrinsic motivation, which does not align with what is shown in the meta in terms of direction or magnitude (*g* = -0.279). The amount of difference between this and other treatments is not statistically significant in the meta in any way; however, there is a small, negative effect that does exist overall. For Unrau and Schlackman (2006), the negative and high magnitude of the effect size in the meta-analysis supports the fact that motivation for reading declined for students over time (*g* = -7.419). It did not support the significant, positive findings reported for intrinsic motivation. The effect size results do support findings within Baker and Wigfield (1999) that show positive and high treatment effects, particularly the fact that high motivation tends to produce high reading achievement and high reading activity (*g* = 11.491). The study applied to students in 5th and 6th grades, and the overall results showed lack of generalizability for the significance that was reported. Not all studies accurately reported the race or ethnicity of the participants, so further research is needed that accurately documents student demographics such as ethnicity, income, and gender. McDonald et al. (2014) had study results that did agree with the meta results in terms of the amount of decrease in inappropriate behaviors exhibited by the students. There was a highly intensive, negative effect size that was calculated for the meta (*g* = -1.897). The reinforcer, though reinforcing to the teachers, did not work as intended for the students. For the students, it mainly functioned as a punisher. The participants were classified as having autism, so the sample for this particular study was not representative of the general population. Although the effect was negative, this outcome was not bad for the students.

Recommendations for Future Research

Reinforcement strategies are varied, and it is a clear choice for differentiation. There is evidence available here and in the meta that shows how varied its applications can be and how it is still useful for students today. It is clear from the data shown, however, that there is a gap that exists between evidence-based practices and scientific, standardized recommendations. The variation of interventions out there, the amount of time given for interventions, as well as the sampling decisions made by teachers and researchers can influence results. Using standards-based practices and having larger samples for reinforcement studies would help to make such projects more generalizable. Having sample size verifications and checks with such tools as GPOWER, Excel, R statistical software, and SPSS can help to tie practice with common research protocols. Educators can talk with researchers, other teachers, and psychologists to determine how to incorporate more reinforcement practices that can span multiple classrooms and use a variety of relevant frameworks. This would help gather more rich data about what is going on and why.

There existed a great amount of reporting bias in terms of methodology and position. Researchers within the literature tended to either show significant gains or losses as a result of incentive use, which exaggerated the practical importance or detriment that statistically was not there when compared to other studies. Even when there were large gains or losses, more context was needed to determine if these results were actually good for all involved. This is not to say that tokens should be discarded as an educational strategy. In fact, the opposite is true. More research needs to be done about them. Statistically speaking, tokens are difficult to generalize. More rigorous approaches other than simple behavior tracking are needed to know what tokens can do for students.

Future research can extend this study and other forms of meta-analyses by creating a review of only meta-analyses on the same topic (Delgado-Rodríguez, 2006). If enough statistical numbers and figures are provided by these metas, then a meta-analysis of meta-analyses could be done as well that would provide more clarity on publication bias and heterogeneity that is more generalizable. Replication studies of old studies that provide more updated, relevant protocols could help to determine if the treatment effects are similar or different to what has already been established within previous studies. Teachers, researchers, and others interested in reinforcement must consider cost, practicality, rigor, and ethics when deciding what would be appropriate for those who would receive tokens as an instructional strategy.

Conclusion

Reinforcement is a strategy that can help address some of the issues seen in schools today, particularly when teachers need a support system that can be adapted to the needs of different students. There was a great amount of heterogeneity found from the analysis of studies. Over 99% was found during data analysis. Because of this, more investigation was needed to find possible reasons that this occurred. Issues with sample size, treatment variation, specific methods used during studies, and reporting were found. Most studies showed positive treatment effects, and some of the results found in the literature could be substantiated by this meta-analysis. The major strength of using a meta-analysis is that it helps to make the results more generalizable and statistically credible. The major limitation is that it cannot correct the methodological concerns or lack of statistical information present within past literature. Not all results could be fully determined or realized, particularly the instances that required ethnicity, gender, socioeconomic status, and special needs status as essential part of analyses. Not everything that was found to be significant within specific studies was actually significant within the broader context; however, there is a wealth of information gleaned that can help with future research into incentives and other areas of behavior analysis. Although the results are relevant to current dynamics in education, it only accounts for the time period of the literature itself. This two-year meta-analysis accounted for noteworthy studies in token reinforcement that were implemented from 1970 to 2017. Other meta-analyses would have to be conducted to determine how newly-created studies would fit in with the previous studies. The findings and recommendations are of interest to teachers, administrators, researchers, and psychologists who would like to increase their strategies and resources for outcomes, specifically ones focusing on performance, behavior, and motivation.

References

Ahn, S., Ames, A., & Myers, N. (2012). A Review of Meta-Analyses in Education: Methodological strengths and weaknesses. *Review of Educational Research, 82*(4), 436-476.

Altman, D.G. (1991). *Practical statistics for medical research* [PDF version]. Retrieved from http://tropical-dendrochronology.org/SHARE/ALTMAN%20(1991)%20-%20Practical%20statistics%20for%20medical%20research.pdf

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct.* Retrieved from http://www.apa.org/ethics/code/principles.pdf

Banks, G., Kepes, S., & Banks, K. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis, 34*(3), 259-277.

Basu, A. (2017). How to conduct meta-analysis: A Basic Tutorial (Unpublished work). *PeerJPreprints 5, 2978*(1), 1-15. Retrieved from https://doi.org/10.7287/peerj.preprints.2978v1

Branch, G., Reid, A., & Plutzer, E. (2021). Teaching evolution in U.S. public middle schools: Results of the first national survey. *Evolution: Education & Outreach, 14*(8), 1-16. https://doi.org/10.1186/s12052-021-00145-z

Bureau, J. S., Howard, J. L., Chong, J. X. Y., & Guay, F. (2022). Pathways to student motivation: A meta-analysis of antecedents of autonomous and controlled motivations. *Review of Educational Research, 92*(1), 46–72. https://doi.org/10.3102/00346543211042426

The Cochrane Collaboration. (2005). *Glossary of terms in The Cochrane Collaboration.* Retrieved from http://aaz.hr/resources/pages/57/7.%20Cochrane%20glossary.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

DeFrancis, Solomon D. (2016). *A qualitative study analysis on how utilizing a token economy impacts behavior and academic success.* (Doctoral dissertation). Brandman Digital Repository. Retrieved from https://digitalcommons.brandman.edu/cgi/viewcontent.cgi?article=1031&context=edd_dissertations

Del Re, A.C. (2015). A practical tutorial on conducting meta-analysis in R. *The Quantitative Methods for Psychology, 11*(1), 37-50. https://doi.org/10.20982/tqmp.11.1.p037

Delgado-Rodríguez, M. (2006). Systematic reviews of meta-analyses: Applications and limitations. *Journal of Epidemiology and Community Health (1979-), 60*(2), 90-92.

Denson, N. & Seltzer, M.H. (2011) Meta-analysis in higher education: An illustrative example using Hierarchical Linear Modeling. Research in Higher Education, 52(3), 215-244. https://doi.org/10.1007/s11162-010-9196-x

Doll, C., McLaughlin, T. F., Barretto, A. (2013). The token economy: A recent review and evaluation. *International Journal of Basic and Applied Science, 2*(1), 131-149. Retrieved from https://pdfs.semanticscholar.org/1870/ad57056432dd3ddb78733879569e213bab13.pdf?_ga=2.150937136.709877635.1569114922-885223096.1569114922

Egger, M., Davey Smith, G., Schneider, M, & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315,* 629-634.

Eliason, B. M., Horner, R. H., May, S. L. (January 2013). Evaluation brief: Out-of-school suspension for minor misbehavior. Retrieved from https://assets-global.website-files.com/5d3725188825e071f1670246/5d8a8b2f88d89eae604e0ea9_EvaluationBrief_130122_revised.pdf

Field, A., Miles, J., Field, Z. (2012). *Discovering statistics using R* (Google Books version). Los Angeles, CA: Sage Publications. Retrieved from https://books.google.com/books?id=Q9GCAgAAQBAJ

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research, 5* (10), 3–8. Retrieved from https://www.academia.edu/285175/Primary_Secondary_and_Meta_Analysis_of_Research

Hackenberg, T. D. (2009). Token reinforcement: A review and analysis. *Journal of the Experimental Analysis of Behavior*, *91*(2), 257–286. https://doi.org/10.1901/jeab.2009.91-257

Hicks, S. (2014). Easy introduction to meta-analyses in R. Retrieved from http://statisticalrecipes.blogspot.com/2014/01/easy-introduction-to-meta-analyses-in-r.html

Higgins, S. (2008). *Using meta-analyses in your literature review* (Presentation). Retrieved from https://www.dur.ac.uk/education/meta-ed/resources/course_material/

Horner, R., Sugai, G., Kincaid, D., George, H., Lewis, T., Eber, L., Barrett, S., Algozzine, B. (July 2012). What does it cost to implement school-wide PBIS? Retrieved from https://assets-global.website-files.com/5d3725188825e071f1670246/5d8a8ca19f7bf86ee571341b_20120802_WhatDoesItCostToImplementSWPBIS.pdf

Houghton Mifflin Harcourt (2019). *5th annual educator confidence report*. Retrieved from https://www.hmhco.com/educator-confidence-report/archived-reports-n

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*(5), 529-554. https://doi.org/10.1016/j.jsp.2011.05.001

Maksimović, J. (2011). The application of meta-analysis in educational research. *Philosophy, Sociology, Psychology and History*, *10*(1), 45-55. Retrieved from http://facta.junis.ni.ac.rs/pas/pas2011/pas2011-05.pdf

Maxwell, J. A. (2012). *Qualitative research design: an interactive approach* (3rd ed.). Thousand Oaks, CA: Sage.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

National Technical Assistance Center on Positive Behavior Interventions and Support. (2017). *Technical guide for alignment of initiatives, programs, practices in school districts.* Retrieved from https://www.pbis.org/resource/technical-guide-for-alignment-of-initiatives-programs-and-practices-in-school-districts

Newman, M. (2003). A pilot systematic review and meta-analysis on the effectiveness of Problem Based Learning. Newcastle: Learning & Teaching Subject Network Centre for

Medicine, Dentistry and Veterinary Medicine. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.6561&rep=rep1&type=pdf

Olejnik, S. F. (1984). Planning educational research. *The Journal of Experimental Education, 53*(1), 40-48. https://doi.org/10.1080/00220973.1984.10806360

Parish, J. G., & Parish, T. S. (1991). Rethinking conditioning strategies: Some tips on how educators can avoid "painting themselves into a corner." *Journal of Instructional Psychology*, *18*(3), 159-166.

Rickert, J. (2014). R and meta-Analysis. Retrieved from https://www.r-bloggers.com/r-and-meta-analysis/

Schmandt-Besserat, D. (1992). *Before writing: Vol. 1. from counting to cuneiform.* [Google Books version]. Retrieved from http://books.google.com/books

Sedgwick, P. (2014). Meta-analysis: Testing for reporting bias. *BMJ: British Medical Journal, 350*, 1-2. https://doi.org/10.1136/bmj.g7857

Shakespeare, S., Peterkin, V.M.S., Bourne, P.A. (2018). A token economy: An approach used for behaviour modifications among disruptive primary school children. *International Journal of Emergency Mental Health and Human Resilience, 20*(2), 1-11. Retrieved from https://www.omicsonline.org/open-access/a-token-economy-an-approach-used-for-behaviour-modifications-among-disruptive-primary-school-children-1522-4821-1000398.pdf

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children*, 31(3), 351-380. Retrieved from https://dropoutprevention.org/wp-content/uploads/2015/05/Simonsen_Fairbanks_Briesch_Myers_Sugai_2008.pdf

Sterne, J., Sutton, A., Ioannidis, J., Terrin, N., Jones, D., Lau, J., . . . Higgins, J. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ: British Medical Journal, 342*(d4002), 1-8. https://doi.org/10.1136/bmj.d4002

Swain-Bradway, J., Lindstrom Johnson, S., Bradshaw, C., & McIntosh, K. (November 2017). What are the economic costs of implementing SWPBIS in comparison to the benefits from reducing suspensions? Retrieved from https://assets-global.website-files.com/5d3725188825e071f1670246/5d76c00cb9339d5f3f267ee7_economiccostsswpbis.pdf

Tan, K. H., Kasiveloo, M., & Abdullah, I. H. (2022). Token economy for sustainable education in the future: A scoping review. *Sustainability, 14*(2), 1-19. https://doi.org/10.3390/su14020716

Tojo, L. M. (2013). Why meta-analysis? A guide through basic steps and common biases. Retrieved from http://blog.efpsa.org/2013/07/15/meta-analysis/

University of Edinburgh, Centre for Cognitive Ageing and Cognitive Epidemiology. (2013). *Systematic reviews and meta-analyses: A step-by-step guide.* Retrieved from https://www.ccace.ed.ac.uk/research/software-resources/systematic-reviews-and-meta-analyses

University of York, Centre for Reviews and Dissemination (2009). Systematic reviews: CRD's guidance for undertaking reviews in health care. Retrieved from https://www.york.ac.uk/crd/SysRev/!SSL!/WebHelp/SysRev3.htm

Weingarten, R. (2019). *The freedom to teach.* Retrieved from the American Federation of Teachers website https://www.aft.org/freedomtoteach

Appendix

Reference List of Studies

(Sorted according to ID number in Forest Plot)

Devers, R., Bradley-Johnson, S., & Johnson, C. M. (1994). The effect of token reinforcement on WISC-R performance for fifth-through ninth-grade American Indians. *Psychological Record, 44*(3), 441-449.

Habaibeh-Sayegh, S. (2014). *The effectiveness of a token economy program in improving behavior and achievement* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 3630049)

Miller, J. B. (1981). *The effects of selected motivational rewards on intelligence test performance of middle school students* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 8128357)

Swain, J. C., & McLaughlin, T. F. (1998). The effects of bonus contingencies in a classwide token program on math accuracy with middle-school students with behavioral disorders. *Behavioral Interventions*, *13*(1), 11-19.

Truchlicka, M., McLaughlin, T. F., & Swain, J. C. (1998). Effects of token reinforcement and response cost on the accuracy of spelling performance with middle-school special education students with behavior disorders. *Behavioral Interventions*, *13*(1), 1-10.

Borrero, C., Vollmer, T. R., Borrero, J. C., Bourret, J. C., Sloman, K. N., Samaha, A. L., & Dallery, J. (2010). Concurrent reinforcement schedules for problem behavior and appropriate behavior: Experimental applications of the matching law. *Journal of the*

Experimental Analysis of Behavior, 93(3), 455-469. https://doi.org/10.1901/jeab.2010.93-455

McDonald, M. E., Reeve, S. A., & Sparacio, E. J. (2014). Using a tactile prompt to increase instructor delivery of behavior-specific praise and token reinforcement and their collateral effects on stereotypic behavior in students with autism spectrum disorders. *Behavioral Development Bulletin*, *19*(1), 40-44.

Hansen, S. D., & Lignugaris/Kraft, B. (2005). Effects of a dependent group contingency on the verbal interactions of middle school students with emotional disturbance. *Behavioral Disorders, 30*(2), 170-184.

Novak, G., & Hammond, J. (1983). Self-reinforcement and descriptive praise in maintaining token economy reading performance. *Journal of Educational Research, 76*(3), 186-189.

Self-Brown, S. R., & Mathews, I. (2003). Effects of classroom structure on student achievement goal orientation. *Journal of Educational Research*, *97*(2), 106-111.

Lynch, A., Theodore, L. A., Bray, M. A., & Kehle, T. J. (2009). A comparison of group-oriented contingencies and randomized reinforcers to improve homework completion and accuracy for students with disabilities. *School Psychology Review, 38*(3), 307-324.

Taylor, D. L. (2000). *The effect of concurrent variable interval reinforcement schedules on children with attention deficit hyperactivity disorder and normal control children* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 9974730)

Marinak, B. A., & Gambrell, L. B. (2008). Intrinsic motivation and rewards: What sustains young children's engagement with text? *Literacy Research and Instruction, 47*(1), 9-26.

Cross, L. M. (1981). *Effects of a token economy program in a continuation school on student behavior and attitudes* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 8124477)

Wulfert, E., Block, J. A., Santa Ana, E., Rodriguez, M. L., & Colsman, M. (2002). Delay of gratification: Impulsive choices and problem behaviors in early and late adolescence. *Journal of Personality*, *70*(4), 533-552.

Hoeltzel, R. C. (1973). *Reading rates and comprehension as affected by single and multiple-ratio schedules of reinforcement within a token economy as measured by precision teaching techniques.* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 74-9066)

Young-Welch, C. (2008). *A mixed-method study utilizing a token economy to shape behavior and increase academic success in urban students* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 3320692)

Popkin, J., & Skinner, C. H. (2003). Enhancing academic performance in a classroom serving students with serious emotional disturbance: Interdependent group contingencies with randomly selected components. *School Psychology Review, 32*(2), 282-295.

Strahan, D. B., & Layell, K. (2006). Connecting caring and action through responsive teaching: How one team accomplished success in a struggling middle school. *The Clearing House, 79*(3), 147-153.

Simon, S. J., Ayllon, T., & Milan, M. A. (1982). Behavioral compensation. *Behavior Modification*, *6*(3), 407-420.

Hayenga, A., & Corpus, J. (2010). Profiles of intrinsic and extrinsic motivations: A person-centered approach to motivation and achievement in middle school. *Motivation & Emotion*, *34*(4), 371-383. https://doi.org/10.1007/s11031-010-9181-x

Unrau, N., & Schlackman, J. (2006). Motivation and its relationship with reading achievement in an urban middle school. *Journal of Educational Research*, *100*(2), 81-101.

Mucherah, W., & Yoder, A. (2008). Motivation for reading and middle school students' performance on standardized testing in reading. *Reading Psychology*, *29*(3), 214-235. https://doi.org/10.1080/02702710801982159

Gaughan, E. J. (1985). *The relationship between point earning behavior and academic achievement in a token economy for emotionally disturbed children* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 8509364)

Abramovich, S., Schunn, C., & Higashi, R. M. (2013). Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educational Technology Research & Development*, *61*(2), 217-232. https://doi.org/10.1007/s11423-013-9289-2

Baker, L., & Wigfield, A. (1999). Dimensions of children's motivation for reading and their relations to reading activity and reading achievement. *Reading Research Quarterly*, *34*(4), 452-477. https://doi.org/10.1598/RRQ.34.4.4

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*(3), 260-267.

McClintic-Gilbert, M., Corpus, J. H., Wormington, S. V., & Haimovitz, K. (2013). The relationships among middle school students' motivational orientations, learning strategies, and academic achievement. *Middle Grades Research Journal, 8*(1), 1-12.

Urdan, T., & Midgley, C. (2003). Changes in the perceived classroom goal structure and pattern of adaptive learning during early adolescence. *Contemporary Educational Psychology*, *28*(4), 524-551. https://doi.org/10.1016/S0361-476X(02)00060-7

Yager, L. (2008). *The relationship between Mississippi school-based rewards programs and the behaviors of 6th grade students.* (Doctoral dissertation). ProQuest Dissertations and Theses database. (Order No. 3358526)

Dreger, K. C. (2017). *Quasi-Experimental study of middle school tokens, behaviors, goals, preferences, and academic achievement* [Doctoral dissertation, Valdosta State University]. Odum Library: Vtext. https://vtext.valdosta.edu/xmlui/handle/10428/2831