ISRG Journal of Arts, Humanities and Social Sciences (ISRGJAHSS)

ISRG PUBLISHERS
Abbreviated Key Title: ISRG J Arts Humanit Soc Sci
ISSN: 2583-7672 (Online)
Journal homepage: https://isrgpublishers.com/isrgjahss
Volume – II Issue-V (September-October) 2024
Frequency: Bimonthly

Developing Software-Based Statistical Models for Educational Incentives in Middle School Classrooms

Kelly C. Dreger, Ed.D.

Valdosta State University

| **Received:** 03.09.2024 | **Accepted:** 09.09.2024 | **Published:** 11.09.2024

***Corresponding author:** Kelly C. Dreger, Ed.D.
Valdosta State University

Abstract

There is a need for more statistical, computerized representations in studies via fixed effects and mixed effects models. This article gives meta-analytic examples of (a) adequate literary, statistical, and conceptual coverage of token reinforcement as defined within educational interventions and (b) practical mixed-effects modeling that is relevant for determining how treatment effect size fits with other characteristics in literature on incentives. The findings from the meta-analytic modeling indicate that sample size, grouping options, timing, study type, and treatment effect size variation have significant influences on the practical significance (effectiveness) of incentives with middle school students. Accounting for these variables helps stakeholders in education develop supports that offer more standardization, versatility, and appeal to students as a whole. A variety of treatment effects for reinforcers may exist, but the overall effect of reinforcement can be positive. This article is recommended for those interested in developing better instructional practices for students, regardless of academic abilities.

Keywords: *Modeling, Reinforcement, Instruction, Education, Supports*

Introduction

Numerous studies about educational incentives describe decision-making frameworks for reinforcement (Doll et al., 2013; Maggin et al., 2011; Schweyer, 2021; Simonsen et al., 2008). Practical strategies for incentivization must be evidence-based in order for teachers and administrators to know what works in the real world. The problem with many incentive-based studies is a lack of in-depth statistical models that generalize, synthesize, and corroborate stated findings. There is a lack of uniformity in terms of program implementation, even though past literature on incentives address the use of them. In education, incentives are frequently stereotyped as tools for struggling learners and individuals, which can influence perceptions and possibilities concerning their use (Dreger, 2017). This creates alarming inconsistencies between what educators say is effective and what the statistical data actually shows. More objective information needs to be available to discover what is actually happening in the classroom and provide a

DOI: 10.5281/zenodo.13744090

way to look at an incentive-based treatment plan in an unbiased way.

A statistical model is one way to objectively represent data while taking into account possible systemic variables and effects. Statistical models help to represent and determine key influences within environments and among participants. Statistical models are important because they can help to account for different factors in the environment, support research claims, explain important phenomena, summarize data trends, and apply STEM (Science, Technology, Engineering, and Mathematics) fields to real-world problems (Gordon, 2019; Winter, 2013). Statistical model construction is not comfortable for everyone, but it can be crucial in demonstrating actual treatment effects and outcomes.

Because of the current state of incentive-based strategies in research and practice, a meta-analytic study was created by the author to address the issues found. The purpose of this article is to determine how statistical, mixed-effects models of reinforcement can be used for meta-analytic research about student support systems. It answers the following questions: (a) To what extent do significant predictors exist for treatment effects pertaining to time and type-based reinforcement with middle school students? and (b) What data trends, if any, exist from meta-analytic, mixed-effects models about incentives for middle school students? The ultimate significance of this research is that it addresses the frequent lack of generalizability, standardization, rigor, and statistical corroboration in educational studies about incentives by demonstrating a system of analysis that reflects reinforcement environments as a whole.

Conceptual and Practical Frameworks

This section provides the theoretical and conceptual foundations necessary for relating important study variables to one another. Empirical and practical frameworks help to make sense of the research questions and the data that are discovered as a result of the research process. Within educational environments, there need to be outcome-based supports in place that ease the burdens, pressures, and problematic situations placed on individuals in the workplace (Maggin et al., 2011; Schweyer, 2022). Incentive systems or any other consequence system must align with decision-making frameworks that are relevant within the school system. This alignment helps synthesize appropriate causes, effects, and goals. Figure 1 provides an illustration of support frameworks that are relevant to reinforcement.

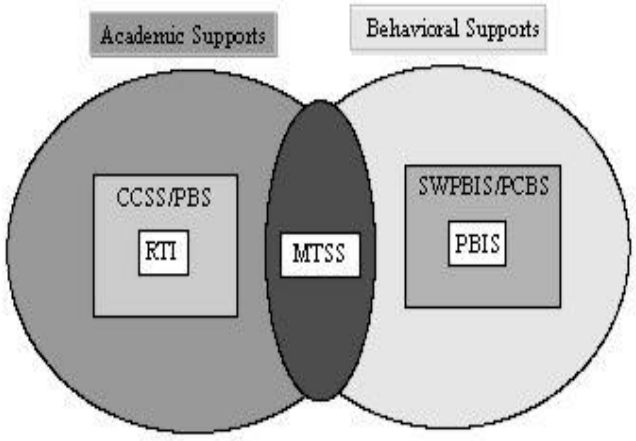

Figure 1. Venn Diagram outlining decision-making frameworks.

Note. CCSS = Common Core State Standards; PBS = Performance-Based Standards; RTI = Response to Intervention; MTSS = Multi-Tiered System of Support; SWPBIS = School-wide Positive Behavior Interventions and Supports; PCBS = Positive Classroom Behaviors and Supports; PBIS = Positive Behavior Interventions and Supports.

An example of an alignment tool that is used for schoolwide incentives with students is found with the National Technical Assistance Center on Positive Behavior Interventions and Support (2017) about Multi-tiered System of Support (MTSS). MTSS addresses tiered supports for achievement and behavior, and it is recommended for K-12 school systems. The systems that help to provide the environmental contexts modeled within the study are as follows: MTSS, Response to Intervention (RTI), Positive Behavior Interventions and Supports (PBIS), Positive Classroom Behaviors and Supports (PCBS), and token reinforcement systems.

Historical Context for Instructional Support Systems

Initial developments for MTSS can be traced back as early as the 1980s because MTSS expands and extends the support systems created within Response to Intervention (RTI) and Positive Behavioral Interventions and Supports (PBIS) (Sugai & Horner, 1999; Sugai & Simonsen, 2012). RTI was a framework that became popular within the Individuals with Disabilities Education Improvement Act (IDEA) in 2004 as an alternative to basing placement decisions solely on intelligence tests and other psychometrics (Fuchs & Fuchs, 2006; Preston et al., 2015). Performance-Based Standards (PBS) were increased during this time to help provide objectives for what students should be able to do each school year. An example of performance-based standards that are used today to shape K-12 curricular needs are the Common Core State Standards (CCSS). RTI supports may be based on guidelines from academic standards. Within RTI, instructional support occurs in three tiers: Tier 1 (general instructional supports), Tier 2 (intensive, small group supports), and Tier 3 (individualized supports). Struggling learners get additional academic supports, particularly with a focus on at-risk students, special needs students, and students with learning disabilities (Preston et al., 2015). PBIS uses a similar 3-tiered system, but the emphasis is on the behavioral needs of students. This is not just about students with disabilities and risks, but it pertains to school-wide, class-wide, and individual supports (Averill & Rinaldi, 2015). MTSS encompasses all supports needed for integrative student success, including academic, social, emotional, and behavioral supports (Averill & Rinaldi, 2015; Howley et al., 2023; Walker et al., 1996). Hill Walker originated the development of MTSS (Walker et al., 1996). Essential initiatives for the MTSS alignment and integration process are as follows: (a) Coordinate and lead alignment process with an executive level team; (b) Define the valued outcome(s) to be achieved; (c) Develop an inventory of the related initiatives that are currently implemented across the district; (d) Identify the core system features for initiatives targeted for alignment; (e) Analyze and make decisions for alignment of initiatives; and (f) Design the plan for effective alignment including implementation, evaluation, and professional development. According to the School Superintendents Association (2014), fidelity is crucial to the success and implementation of supports systems, particularly the Positive Behavioral Interventions and Supports (PBIS) framework. PBIS and similar support systems address intervention requirements that are in compliance with the Individuals with Disabilities Education Act (von Ravensberg & Blakely, 2017). These expectations are useful for RTI and MTSS frameworks as well.

DOI: 10.5281/zenodo.13744090

Positive Classroom Behavior Support

Furthermore, Swain-Bradway et al. (2017a, 2017b) identify Positive Classroom Behavior Support (PCBS), which is different from PBIS. PCBS is an umbrella term for positive practices in relation to behavior, whereas PBIS is a type of tiered system that is an example of PCBS. When PBIS is incorporated throughout schools, it is often known as SWPBIS. SWPBIS would be on the same level as PCBS because it meets similar requirements. This brings up the fact that PBIS is not always implemented for all districts, schools, or classrooms. The program breadth and depth can vary according to school system. MTSS would also be an example of a system that has PCBS. PCBS data can address the areas of fidelity, outcomes, and equity in system practices. Interventions may address various areas, including settings, routines, expectations, prevention strategies, responsive strategies, organizational funding, and data systems (Siegel, 2021; Swain-Bradway et al., 2017a, 2017b; U.S. Department of Education, 2021). A fully-functional PCBS system needs the following before and during implementation: (a) PCBS training, (b) School-wide positive behavioral support practices, (c) Policies and operating procedures for recruiting and hiring staff, (d) Clearly defined policies and procedures, (e) Ongoing professional development opportunities, (f) School investment in evidence-based curriculum, (g) Investment in district-wide data systems, (h) Collection and use of classroom data for decision-making, (i) Specific school-wide strategies for positivity, reinforcement, and expectations, (j) A formal process exists for requesting assistance, and (k) Policies and operating procedures for annual evaluation of personnel and systemic features. Based on the information in this list, it is observed that reinforcers are used within and alongside other types of systems.

Classroom Reinforcement

Additionally, Simonsen et al. (2008) discovered essential characteristics of effective, evidence-based classroom management, which included reinforcement: (a) Maximize structure; (b) Post, teach, review, monitor, and reinforce expectations; (c) Actively engage students in observable ways; (d) Use a continuum of strategies for responding to appropriate behaviors; (e) Use a continuum of strategies to respond to inappropriate behaviors. They go further to say that reinforcement is used as a strategy for responding to both appropriate behaviors and inappropriate behaviors. Reinforcement refers to an event-based consequence that increases the likelihood for a behavior to occur again (Ferster & Skinner, 1957). It can improve engagement, social skills, motivation, behavior, and self-reflection during class activities. For instance, it is possible to have tokens address particular groups of students, such as those with autism (Whitney et al., 2018). A basic definition of tokens that is often used within research is provided by Skinner (1953). A token is a "generalized reinforcer distinguished by its physical specifications" (Skinner, 1953, p. 79). In other words, a token is an item with physical properties that is used as an exchange system for goods and services. Examples include money, points, tickets, badges, coins, coupons, stars, stickers, and checks. In a token system for the school setting, students accumulate items for appropriate behaviors, actions, and situations. These items are exchanged for more desirable items. Prize lists are developed for activities that require a token exchange. Token reinforcement systems can be schoolwide, classwide, and individuated. Token options "can depend on the setting, population, manager's or teacher's preference, cost, among other considerations" (Doll et al., 2013, p. 134). Tokens can be included with students of varying abilities, but the reinforcement used may not necessarily reflect what is preferred by all students. The amount of choice and the structure of token systems would determine their feasibility. The idea of token reinforcement, however, does not imply mandatory ability grouping. Ability grouping is a concept required within merit systems, but it is optional within token systems.

Summary of System Connections

A multi-tiered approach to educational intervention provides an alternative in decision-making that can be used separately from and simultaneously with standardized measures of intervention. MTSS instructional supports combine tiered supports found in RTI and PBIS. Effective MTSS in schools have PCBS as part of their implementation. Reinforcement is a requirement of PCBS, and an example of a reinforcement system is a token system. This means that MTSS, PBIS, and RTI can include token systems as an option for instructional support because token systems can address situations related to behavior, performance, social communication, emotional regulation, and motivation.

How do the intervention systems address the research questions? All of the data used for this study have forms of practical reinforcement for students. The first research question is asked to provide clarity on the types of factors that influence the effects of token reinforcement systems used for educational purposes. If significant predictors can be found, then those who implement and design support systems such as MTSS can be better informed about what consistently works for students. Teachers and administrators can help create more effective learning environments when they are aware of what is happening. Their decisions would be based on outcomes that are more representative of students as a whole. The second research question is asked in order to determine if there are any statistical trends in the data. Token systems are sometimes perceived as relevant only to struggling learners or traditional psychologists. It is important to know if this perception actually matches the factual results found from a variety of studies about different reinforcement systems.

How is a meta-analytic study useful in this education context? A meta-analysis is a study of studies, encompassing a large body of research from multiple sources about a particular topic of interest (O'Rourke, 2007; Paul & Barari, 2022). It is more thorough than a literature review, and it includes statistical information to support literary and empirical claims. It can be used to further support claims in systematic reviews, research studies, and educational practice. It is specifically used to discover trends in data, increase statistical power of claims, and to determine more generalized, unbiased results. Token incentives are known as a strategy of differentiation, but there are other factors that can influence the results. For instance, results from a mixed methods study in Dreger (2017) indicated performance outcomes were influenced by ability grouping. A quasi-experiment, interviews, and focus groups were conducted for the study. The control group had higher math scores than those who received token interventions (i.e., points and coins). The control group contained more accelerated students than the treatment groups as well. This means that more struggling students received the token interventions from the teachers, which was not discovered until the interviews were conducted. Two out of the three math teachers directly interviewed about token interventions had the perception that struggling students needed the interventions more, which could have further influenced the administration of the treatment. The meta-analysis in this article was conducted to

DOI: 10.5281/zenodo.13744090

determine (a) if performance-based and ability-based influences extended to other studies in education and (b) if there were other influences that needed to be discovered. The two research questions address these areas of concern with modeling in mind. Because modeling and testing were conducted in R statistical software to answer the research questions, the tests generally assume that the null hypothesis is true. This is our hypothesis for the meta-analysis, which means we expect no significant differences between and among the groups analyzed.

Methods

Permission was obtained to conduct a quantitative meta-analysis in 2018 as a follow-up to the results of Dreger (2017), which was supervised by the Institutional Review Board (IRB) at Valdosta State University. Figure 2 illustrates the process required to conduct this study. After approval was obtained, studies found to be relevant to reinforcement were summarized. The studies within this article pertain to token reinforcement use within middle grades as an instructional intervention. The focus on middle grades and token reinforcement was specifically done to build on previous mixed methods research, descriptive research, and analytic research by the author (Dreger, 2017; Dreger & Downey, 2021; Dreger & Downey, 2022). Dreger & Downey (2022) also address the same meta-analysis discussed in this article; however, this article expands on the previous discussion by providing conceptual and statistical information not covered within the previous article. Research databases within EBSCOhost, ProQuest, and PsycINFO were used to acquire past studies pertaining to token reinforcement and middle grades information. Search terms paired with the words reinforcement and middle grades included type (i.e., schedule, curriculum, students, design, teacher, classroom, environment, instructional strategies, and testing), student development, goals, socioeconomic status, culture, race, gender, attendance, attention, politics, location, parental involvement, and community partnerships.

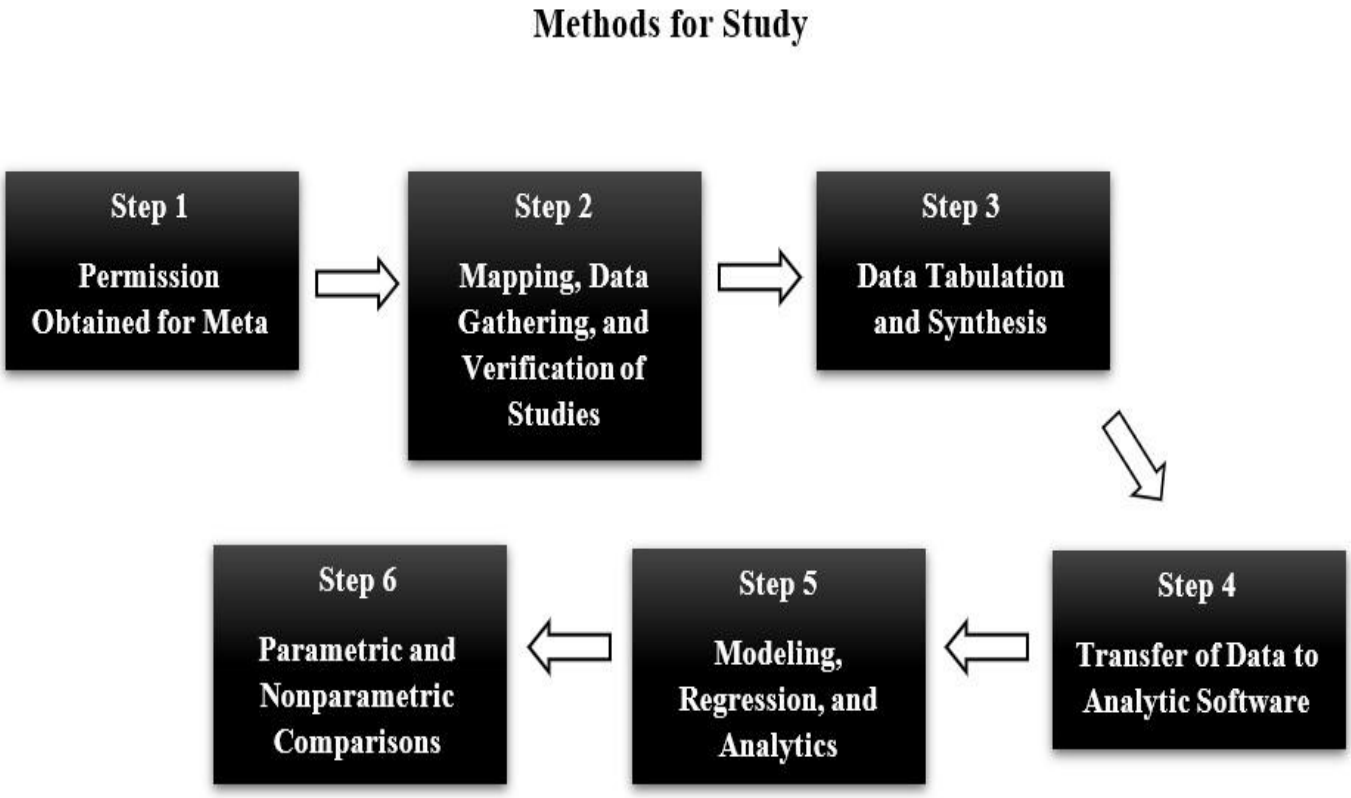

Figure 2. General Methodology for Study

To document validity and reliability of each study, a checklist was created from recommendations made by Creswell (2009), Creswell and Plano Clark (2011), Maxwell (2012), and Ahn et al. (2012). Each of the studies were assessed according to items on the checklist (See Appendix A). The studies were given a grade based on the amount of items that existed within the given studies. There were 32 items on the checklist. A score of 23 or above was required in order for studies to have a passing score (i.e., at least 70%).

Out of the 129 studies reviewed, there were 31 studies from journal articles and dissertations that had enough satisfactory information for meta-analyses or effects-based modeling (n = 5,765). Most of the studies contained the information necessary for basic procedures. The basic information required for data analysis are means and standard deviations for a particular dataset. These means and standard deviations were provided for the study by default within previous literature, or they were calculated from the raw data given within previous studies for this topic. In Table 1, important characteristics of each study (i.e., sample size, year, location, study type, outcome measured, and grouping variable) are summarized in chronological order.

DOI: 10.5281/zenodo.13744090

Table 1
Important Study Characteristics

Study	*n*	Year	Location	Type	Outcome	PGV
Hoeltzel	4	1973	West	EC	Performance	Time
Cross	86	1981	Other	EC	Behavior Performance	Time
Miller	135	1981,1985	South	EC	Performance	Treatment
Simon, Ayllon, and Milan	7	1982	South	CA	Performance Behavior	Treatment
Novak and Hammond	28	1983, 2001	West	EC	Performance	Treatment
Gaughan	40	1985	Northeast	CA	Performance	Time
Ames and Archer	176	1988	Midwest	SQ	Motivation	Time
Devers, Bradley- Johnson, and Johnson	25	1994	Midwest	EC	Performance	Treatment
Truchlicka, McLaughlin, and Swain	3	1998	West	EC	Performance	Time
Swain and McLaughlin	4	1998	West	EC	Performance	Time
Baker and Wigfield	370	1999	South	SQ	Motivation Performance	Time
Taylor	60	2000	West	EC	Performance Behavior	Treatment
Wulfert	114	2002	Northeast	EC	Performance Behavior	Treatment
Popkin and Skinner	5	2003	South	CA	Performance	Time
Urdan and Midgley	555	2003	Midwest	SQ	Performance	Time
Self-Brown and Mathews	71	2003	South	EC	Motivation	Treatment
Hansen and Lignugaris/Kraft	9	2005	West	EC	Behavior	Time
Strahan and Layell	479	2006	South	CA	Performance	Treatment
Unrau and Schlackman	470	2006	West	CA	Performance	Time
Marinak and Gambrell	75	2008	Northeast	EC	Motivation	Treatment
Young-Welch	400	2008	Midwest	EC	Behavior Performance	Treatment
Mucherah and Yoder	388	2008	Midwest	CA	Performance	Time
Yager	60	2008	South	SQ	Behavior	Treatment
Lynch et al.	6	2009	Northeast	EC	Performance	Time
Borrero et al.	3	2010	Other	EC	Behavior	Time
Hayenga and Corpus	343	2010	West	CA	Performance	Time
Abramovich, Schunn, and Higashi	51	2013	Northeast	SQ	Motivation	Time
McClintic-Gilbert et al.	90	2013	West	SQ	Motivation	Time
Habaibeh-Sayegh	60	2014	Midwest	EC	Behavior Performance	Time
McDonald et al.	3	2014	South	EC	Behavior	Time
Dreger	205	2017	South	EC	Performance	Treatment

Note. EC = Experimental/Causal; CA = Correlation/Ambiguous Methods; SQ = Survey/Questionnaire; PGV = Primary Grouping Variable.

DOI: 10.5281/zenodo.13744090

Microsoft Excel spreadsheets were made to transfer key information into data tables. Data tables with appropriate formatting for statistical analysis were transferred to R Statistical Language (Base Software), RStudio, and R Commander to start and complete the data analysis stage of the information gathered. The tabular data were used to create meta-analytic models of the 31 studies. Figure 3 summarizes all of the parametric and non-parametric models necessary for the study.

Model Diagram for Testing and Analysis (3 Levels)

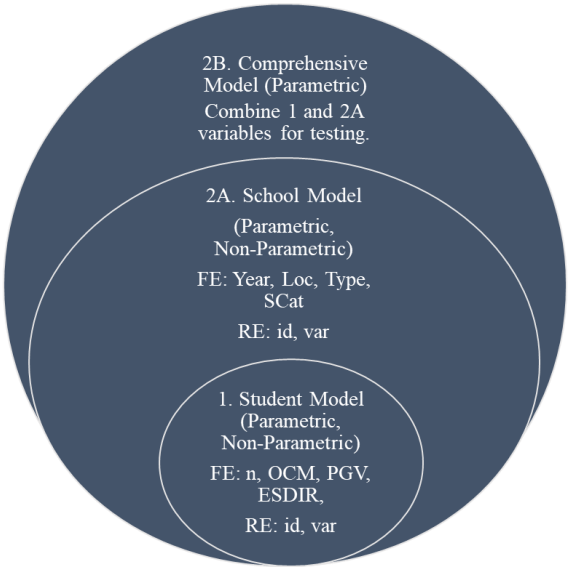

Figure 3. Diagram of parametric and non-parametric modeling for this study, where FE = Fixed Effects, RE = Random Effects, Year = Study Year; Location = Study Location; Type = Study Type; SCat = Number of Sessions; id = Study id; var = Variance; n = Sample size; OCM = Study Outcome; PGV = Primary Grouping Variable; ESDIR = Effect Size Direction.

The Level 1 Model is the individual, student level. The Level 2 Model A equation is the school level, general study characteristics. Level 2 Model B is a comprehensive model that includes all relevant variables from both levels. All models have effect size as the dependent response variable. The specific procedures that were necessary to complete the modeling were as follows: Gather Data, Create the Model(s), Test Assumptions, Test the Model(s), and Interpret the Data. A checklist was developed in order to keep track of the strategies needed to complete each step (See Appendix B). Due to the fact that the school environment contains both expected and unexpected occurrences that may affect treatment effects, mixed effects modeling was the most appropriate method of modeling for this investigation. Mixed effects modeling is similar to basic regression in that there are predictors, response variables, and one or more equations based on specific effects; however, there are key differences in terms of accountability and complexity (Dreger & Downey, 2022; Gordon, 2019). In this instance, there are multiple levels of modeling that reflect the multiple tiers of supports (i.e., individual and group) that are necessary in such systems as MTSS, PBIS, and RTI. Variables necessary for modeling were the following:

- Effect Size (es) – It "refers to the magnitude of the relation between the independent and dependent variables, and it is separable from statistical significance, as a highly significant finding could correspond to a small effect , and vice versa, depending on the study's sample size" (Funder & Ozer, 2019, p. 156). Factor-based thresholds of Small, Medium, Large, and Trivial were created during rank-based analysis. Hedges' g was used to calculate effect size.
- Sample Size (*n*) – The number of participants in a study sample, symbolized as *n*. Like effect size, initial testing involved a continuous variable until rank-based modeling had to be performed. Factor-based thresholds of Small, Medium, Large, and Trivial were created during rank-based analysis.
- Outcome (OCM) – Continuous, numerical results for performance, behavior, and motivation that were the basis for a study. For instance, performance scores would be an outcome for a study about student performance differences. This is symbolized as OCM.
- Primary Grouping Variable (PGV) – Pertains to how study treatments were organized. In terms of reinforcement, there are type-based and time-based studies. Coded as 1 (time) and 2 (type).
- Effect Size Direction (ESDIR) – The sign of an effect size number that tells the direction of the effects (positive or negative). Coded as 1 (positive) and 2 (negative).
- Identification number (id) – The number assigned to each study to keep track of what was read. The first work of literature read received the first id number.
- Variance (var) – Effect size variability and dispersion, represented as a number. This was used as a continuous variable until ranked-based analysis, where explained variance was categorized into Trivial, Small, Medium, and Large.
- Study Year (Year) – The year in which the study was completed, sorted according to important literary periods found from the literature review: (a) 70s to 80s, (b) 90s to 00s, and (c) 00s to 10s.
- Location (Loc) – Where the study took place and/or the affiliated areas pertaining to study implementation. Categories used were Northeast, Midwest, South, West, and Other. The defined categories pertain to studies located in the United States. Other refers to

DOI: 10.5281/zenodo.13744090

ambiguous locations and combinations of locations that do not specify a particular region in the United States.

- Study Type (Type) – The design of the study (i.e. experimental, ambiguous yet correlational, survey/questionnaire).
- Number of Sessions (SCat) – The number of sessions required to complete the study, sorted according to frequency (At Most Two, Multiple Sessions, Weeks to Months, Year at Least).
- Error (*e*) – Representative of anything not accounted for within the study that would be of concern, including programming errors, unknown confounding, reporting bias, technical glitches and statistical mistakes. In mixed-effects modeling, this would not be numerically calculated.

After general models were chosen that fit the data, specific procedures for model fitting and analysis could start. The modeling procedures in R were recommended by Christensen (2016, 2019), Del Re (2015), Koller (2016), Kuznetsova et al. (2017), McNeish and Kelley (2018), Lawson (1983), Mertler and Vannatta (2013), Tabachnick & Fidell (2007), and Winter (2013). There were six essential assumptions that were tested within the data variables: linearity, collinearity, independence, influential weights, equal variances, and normality. Then, the actual models were tested from this point. Initial testing involved the parametric, untransformed WB-MEM models in Figure 2. They were named Model Level 1, Model Level 2A, and Model Level 2B. They were stored in R as follows:

- Level 1 Model: es ~ *n* + OCM + PGV + ESDIR + (1|id) + (1|var) + *e*
- Level 2 Model A: es ~ Year + Loc + Type + SCat + (1|id) + (1|var) + *e*
- Level 2 Model B: es ~ Year + Loc + Type + SCat + *n* + OCM + PGV + ESDIR + (1|id) + (1|var) + *e*, where es = effect size, Year = study year, Loc = location, SCat = number of sessions, *n* = sample size, OCM = outcomes, PGV = primary grouping variable, ESDIR = effect size direction, (1|id) = id number in random effects form, (1|var) = effect size variance in random effects form, *e* = error

After transformations and initial testing, the models were labeled non-parametric and non-linear due to violations within the results. Once non-linearity was fully established, supplemental models were needed to explain additional violations within the untransformed models. The WB-MEM Models were transformed into ordinal models in Clmm2:

- Level 1, Model 1: es ~ *n* + OCM + PGV + ESDIR + (1 | id) + *e*
- Level 1, Model 2: es ~ *n* + OCM + PGV + ESDIR + (1 | var) + *e*
- Level 2, Model 1: es ~ Year + Loc + Type + SCat + (1 | id) + *e*
- Level 2, Model 2: es ~ Year + Loc + Type + SCat + (1 | var) + *e*

Ordinal procedures are appropriate when parametric assumptions are violated or there are bivariate/ordinal data that exists in the dataset (Cangur et al., 2018). ANOVA-based likelihood ratio tests and ranked ANOVAs (RANCOVAs) with smoothing functions were completed using the four Clmm2 ordinal models. The packages car, compute.es, effects, ggplot2, multcomp, and WRS2 were used to analyze the data. The RANCOVA procedures were recommended by McSweeney and Porter (1971). Olejnik and Algina (1984, 1985) point out that these procedures are modeled after and have similar results to the Quade (1967) method.

Results

Statistical results for the 31 studies are explained individually and in aggregate. Relevant subgroups and percentages are presented in Table 2. The random effects id (id number) and var (effect size variance) did not have subgroups, so they were not included in Table 2.

Table 2

Counts and Percentages for response and fixed effects

Variable	Subgroup for Variable	Study Count	Study Percentage
ES	Trivial	9	29.03
ES	Small	6	19.35
ES	Medium	4	12.90
ES	Large	12	38.71
n	Trivial	9	29.03
n	Small	8	25.81
n	Medium	12	38.71
n	Large	2	6.45
OCM	Performance	15	48.39
OCM	Behavior	4	12.90
OCM	Motivation	5	16.13
OCM	Combination	7	22.58
PGV	Time	18	58.06

DOI: 10.5281/zenodo.13744090

PGV	Treatment	13	41.94
ESDIR	Positive	19	61.29
ESDIR	Negative	12	38.71
Year	70s-80s	7	22.58
Year	90s-00s	17	54.84
Year	00s-10s	7	22.58
Loc	Northeast	5	16.13
Loc	Midwest	6	19.35
Loc	South	9	29.03
Loc	West	9	29.03
Loc	Other	2	6.45
Type	Experimental/Causal	18	58.06
Type	Correlation/Any Relationship	7	22.58
Type	Survey/Questionnaire	6	19.35
SCat	At Most Two	6	19.35
SCat	Multiple Sessions	8	25.81
SCat	Weeks to Months	12	38.71
SCat	Year at Least	5	16.13

Note. Relevant counts and percentages for analyzed variables. ES = Effect size; *n* = sample size; OCM = Outcome Type; PGV = Primary grouping variable; ESDIR = Effect size direction; Year = Study Year; Loc = Location; Type = Study Type; SCat = Categories for Number of Sessions.

Based on the statistical information found from the models, did significant predictors exist within the models? Significant predictors did exist within the models. There were a total of 48 influential points during linear model testing for beta weights. Values from Yager (2008), Unrau and Schlackman (2006), Baker and Wigfield (1999), and Devers et al. (1994), were frequently found within the influential points. The Clmm2 models produced significant results as well. For the first model in Level 1, there were significant results found in terms of sample size and effect size within the large category ($p < .001$). For the second model, significant effects were shown within sample size, but this time the trend was found among all ranks ($p < .001$). The Level 2 models showed no significant estimates of effects within their results ($p > .05$).

From the results within Level 1 RANCOVA modeling, it was determined that time had an influence on effect size when paired with sample size as a covariate ($p = .042$). The variables of time and type of groups (independent or dependent) had an influence on effect size when accounting for direction as a covariate ($p < .001$). This was also true when accounting for variation ($p = .004$ for group; $p = .002$ for time). Through principal component analysis with the raw data as recommended by Hayden (2018) and Winter (2013), it was determined that sample size accounted for 72.2% of the variation and effect size variance accounted for 27.8% of the variation.

Were there data trends that stood out within the data? There were unique data trends and unique inconsistencies. The assumption of linearity was not met for the mixed-effects models before or after transformations. Residual results and plots from the R packages indicated patterns inconsistent with normal and linear distributions. For non-parametric y ~ x correlation testing, the results of both Spearman Rho and Kendall Tau showed significant correlations between pairs: a) sample size and study type, b) id and study type, c) sample size and time sessions, d) sample size and effect size variance, e) id and sample size, f) effect size direction and id, g) effect size direction and variance, and h) id and effect size variance. Significant correlations are shown in Table 3.

Table 3

Significant correlations between fixed and random effects

x (Variable 1)	y (Variable 2)	Spearman Rho (coefficient)	Spearman Rho (*p*-value)	Kendal Tau (coefficient)	Kendall Tau (*p*-value)
n	Type	- .419	.019	-.332	.023
id	Type	- .795	< .001	-.698	<.001
n	SCat	.376	.037	.296	.035

DOI: 10.5281/zenodo.13744090

n	var	.767	< .001	.581	<.001
id	*n*	.553	.001	.399	.002
ESDIR	id	.400	.026	.332	.029
ESDIR	var	.429	.016	.356	.019
id	var	.421	.019	.308	.015

Note. *n* = sample size (fixed effect); Type = study type (fixed effect); id = identification number (random effect); SCat = time sessions (fixed effect); var = effect size variance (random effect); ESDIR = effect size direction (fixed effect).

Out of the eight correlations found, four correlations involved sample size. Four involved id number as well, which was given after studies were reread for sorting purposes. The lowest significance values were found between a) id and study type and b) sample size and variance. Bivariate normality was tested using Shapiro-Wilk (SW) and Kolmogorov-Smirnov (KS) within the EZR package and RCommander. Residual structures for each variable indicated non-normality and nonlinearity except in the case of the id variable, which had a $p = .247$ on the initial SW test for normality. The p value increased to $p = .996$ using the KS test.

Multivariate normality was tested using the MVN package. Royston and Mardia were the recommended tests. The testing had three sections: multivariate normality, bivariate normality, and descriptives. For Mardia, only partial requirements were met for some of the pairs. Overall, many of the pairs did not show normality ($p < .05$). Royston's tests indicated that no pair had multivariate normality ($p < .05$). Both tests did conclude that id had normality by itself ($p > .05$), which was also found within the bivariate tests already performed on the variable. For the effect size (Hedges' *g*), the top nine studies had an effect size over one. This result is statistically possible but rare in terms of effects size. Table 4 shows the top 11 effect sizes, including the previous study done by Dreger (2017).

Table 4

Top 11 Means, Standard Deviations, and Effect Sizes from Effects Modeling, Sorted by Effect Size Magnitude (High to Low)

Study	*CM*	*CSD*	*EM*	*ESD*	*g*	*Vg*
Yager (2008)	.16	.27	.50	.16	18.505	1.038
Baker and Wigfield (1999)	2.85	.26	25.94	28.61	11.491	.166
Unrau and Schlackman (2006)	2.82	.01	2.71	.01	-7.419	.034
Novak and Hammond (1983)	2.34	.00	7.33	1.44	3.619	.549
McDonald et al. (2014)	23.80	4.50	11.93	5.46	-1.897	.727
Popkin and Skinner (2003)	46.76	31.13	54.12	37.66	1.796	.306
Hansen and Lignugaris/Kraft (2005)	.12	.14	.31	.03	1.772	.289
Lynch et al. (2009)	73.67	.00	91.75	1.90	1.671	.254
Swain and McLaughlin (1998)	54.50	22.52	83.00	2.16	1.549	.528
Devers and Bradley-Johnson (1994)	94.03	5.30	103.10	6.87	.856	.105
Dreger (2017)	79.76	2.07	65.94	3.71	-.843	.017

Note. *CM* = Control Group Mean; *CSD* = Control Group Standard Deviation; *EM* = Experimental Group Mean; *ESD* = Experimental Group Standard Deviation; *g* = Hedges' g (Effect size); *Vg* = Effect Size Variation

Treatment effects and traits found in (a) Yager (2008), (b) Unrau and Schlackman (2006), (c) Baker and Wigfield (1999), and (d) Devers et al. (1994) carried more weight than the other 27 studies. Three out of the four studies had performance-based reinforcement as a focus. When adding all the 31 effect sizes together, the sum of the effects is 33.28, and most of that can be found within Yager (2008) and Baker and Wigfield (1999) alone. Just adding those two effect sizes together gives a total of 29.996. What both of the studies had in common were the location and the study type.

Discussion

Based on the information found in the first research question, the researcher found that the amount of influence depended on sample size and approach. A high, positive treatment effect was found as a whole; however, what works for students in individual classrooms may not show up as having much effect in the general sense. To get a better idea of what can work, more details have to be given about specific studies. The literature from Yager (2008), Baker and Wigfield (1999), Unrau and Schlackman (2006) and others such as Schweyer (2022) and Swain-Bradway (2017a, 2017b) indicate that positive effects from token reinforcement are possible, especially when they are based on the needs and goals of students. Having goals, standards, and objectives that are academic in nature, whether they are used for short-term performance or long-term mastery of skills, not only are an important focus for educators, but they also are reasonable components to include when developing instructional interventions about reinforcement. Not all studies had positive effect sizes, but the ones that did often cited generalizability as a limitation. The top two effect sizes were found with Yager (2008) and Baker and Wigfield (1999), but both had

DOI: 10.5281/zenodo.13744090

issues with generalizing the results beyond their participant pools (Dreger & Downey, 2022). Both were major influences in the study model, but they were not true experiments or quasi-experiments. Yager (2008) contained survey research for middle school students in Mississippi. Baker and Wigfield (1999) included a variety of case-based assessments in their analyses as well as the administration of the Motivation for Reading Questionnaire. Both were studies that involved surveys or questionnaires. They did not focus on the same outcomes. Yager (2008) focused on school-wide behavioral reinforcement (n = 60), and Baker and Wigfield (1999) encouraged the use of performance-based incentives and motivation-based incentives for groups of students (n = 370). Out of the 31 studies, there were 18 (58.06%) that were experimental, quasi-experimental, or causal-comparative in nature. Only six (19.35%) used surveys or questionnaires.

The results from the second research question indicated that accurate models based on MTSS, PBIS, RTI and similar frameworks did not follow what is typically expected of the general population. Parametric modeling assumes a bell-curve, also known as a normal distribution. In practical terms, it is unrealistic to expect this when dealing with student support systems. If the goal is for all students to meet or exceed the standards and expectations that are placed on them, then the model for that would have to be non-parametric on one or more levels, with allowances for both fixed and random occurrences. Some support systems, such as the ones found in Gaughan (1985), Hansen and Lignugaris/Kraft (2005), Lynch et al. (2009), McDonald et al. (2014), and Popkin and Skinner (2003), only address students with atypical behaviors and needs. All had positive treatment effects except for McDonald et al. (2014), which focused on behavior-based incentives. Hansen and Lignugaris/Kraft (2005) also focused on behavior, but the remaining three studies encouraged the use of performance-based incentives that were based on the participants' needs and preferences. Performance-based incentives tended to show more positive effects overall than other incentive types (i.e., behavior, motivation, and combination). The findings revealed that token use was not automatically detrimental to targeted outcomes. The drawback is that when looking at results from an overall standpoint, it is hard to determine what should be kept and what should be discarded in order for token reinforcement to have long-lasting, game-changing effects that work for middle school students. When it comes to model selection and model fitting, being able to determine non-normal distributions, specific non-parametric variables, and statistical software for handling such data helps to show how practical studies are and how they actually differ from ideal expectations of what researchers would like to see (Foldnes & Olsson, 2016; Gu & Ma, 2005; Noguchi, Gel, Brunner, & Konietschke, 2012; Tremblay & Newman, 2015; Tsangari & Akritas, 2001; Zuur, Ieno, Walker, Saveliev, & Smith, 2009).

In terms of the hypothesis, the results did not show what was expected by the researcher. All levels showed non-normal results with high heterogeneity, and they had to be transformed in order to be properly fitted with the appropriate models. Upon further inspection within the transformed models, significant predictors existed with student-level variables (Level 1) and not with school-level variables (Level 2). Therefore, all models showed that the token systems were mostly influenced by students' needs and the classroom environment at the time of instruction. There were 22 studies that had performance as an outcome, but outcome was not found to be a significant influence on effect size for the studies on either level. What was a major influence was time, specifically whether or not groups were time-based in nature. The top 11 studies, when sorted by magnitude, mostly had time-based designs. This means that treatment strength tended to show up higher with time-based designs. Time-based designs in this circumstance usually have dependent grouping, where the same group of participants are observed before and after a specific type of reinforcement. This differs from treatment-based designs, where independent groups receive different treatments over a similar period of time. Within the study data, time-based designs tended to have small groups. It is important to note that in this study, larger sample sizes tended to have higher positive effects, but smaller sample sizes tended to have higher magnitude. How is this possible? There are two explanations for this: (a) Some studies carry more statistical weight and influence than others in the calculations, and (b) Effect size direction is not the same as magnitude. Based on the data, the most appropriate reinforcement design that is likely to produce strong results is one with a time-based design. Although Yager (2008) is a treatment based design, there are seven studies in the top 11 that are time-based. Treatment-based designs and time-based designs with medium or large sample sizes are likely to have strong influence and positive effects. Baker and Wigfield (1999), Devers et al. (1994), Unrau and Schlackman (2006), and Yager (2008) had the most influence within the calculations. Two had time-based designs and two had treatment-based designs. Out of these four, sample size ranged from 25 participants for Devers et al. (1994) to 470 participants for Unrau and Schlackman (2006). The study year and length of sessions are separate variables with no significant predictive power. Therefore, educators and researchers need to pay more attention to treatment timing according to instructional groups and activity changes within those groups. This supports the ability group influence found within Dreger (2017) since teachers design their lessons in ways that take into account the amount of time they want to spend on instructional activities and how students need to be grouped for instructional support. The meta-analysis also supports the idea that different systems can produce different results, and reinforcement systems can be successful with students who have different learning needs, behaviors, and personalities. Teachers' perceptions and expectations may not always align with statistical results.

There are various implications that exist because of the modeling procedures completed within this article. Firstly, it is possible to model incentive effects and probable influencers in a way that is relevant to the school system today. Similar to the structure of the meta-analytic models discussed, teachers and researchers can propose one overarching intervention support system that works with different incentive systems. Each incentive-based subsystem can be implemented simultaneously within the one system. MTSS can be utilized for this arrangement because it facilitates the use of a tiered support system with a variety of supports in place for students. Secondly, it is possible to have positive treatment effects for token interventions as a representative whole, even if not all studies support the positive finding. The supports included in a reinforcement system could be arranged or changed according to actual evidence-based results; however, it is important to allow for diverse systems without compromising practical integrity. Thirdly, decision-making about intervention use can be more tailored to students' needs since it is known statistically where strengths and weaknesses exist within the educational studies discussed. Heterogeneity, sample size, variation, and instructional timing are all important factors to consider during implementation of token-

DOI: 10.5281/zenodo.13744090

based incentives. Future meta-analytic models can and should be created with these and other variables in mind.

Recommendations

The reason mixed-effects modeling is important in research and practice is because it represents situations that have clear influencers (fixed effects) and situations that have unpredictable influencers (random effects). This is needed if educators, researchers, and individuals want to determine possible explanations for outcomes in a way that is actually supported by factual statistics. When explaining an evidence-based treatment that could be adopted into schools and businesses, there needs to be a determination of why it is important on a larger scale or how it would need to be modified to work with different types of people who come from diverse backgrounds.

With this in mind, one recommendation is to use sampling techniques that reflect the typical classroom setting in the region of the study. Instead of using reinforcers just for struggling students, educators can identify a subset of students that are representative of different types of learners. There are too many studies within reinforcement that have limitations in terms of generalizability. Using frameworks such as RTI, PBIS, and MTSS to inform decisions would help to make supports more accessible for different learners. Encouraging more social, electronic, and network-based reinforcement in studies in addition to what is available out there helps modernize the practice of reinforcement for all those involved. Having participant groups primarily based on treatment and not age or ability helps to mitigate undue influences seen with time-based scheduling. There will be elements of time present during schooling, but the amount of focus given to time is something that needs to be planned in advance when discussing instructional supports.

Reinforcement systems are usually developed as a way to regulate past behavior, but they are also used in order to encourage future behavior. Reinforcement systems, in other words, are not just feedback systems. They are feed-forward systems as well. According to Andreas (2012), feed-forward systems are known as 'strategic planning' or 'backward planning' (p. 41). Systems that are goal based and outcome based would be feed-forward in nature. More research and instructional supports should include the concept of feed-forward systems so that future studies on reinforcement can determine essential differences in the feedback and feed-forward processes.

Future research also needs to be available that gives more statistical depth to projects on motivation. Many traditional studies have been experiments or surveys with simple behavior tracking. More complex and mixed methodologies that involve parametric modeling, non-parametric modeling, observations, interobserver agreement, member checking, reliability statistics, triangulation, and interviews are strongly encouraged. Researchers who use modeling effectively can generate descriptives for central tendency, descriptives for dispersion, fixed effects modeling, mixed effects modeling, random effects modeling, or some combination of models.

Conclusion

Mixed-effects modeling can represent outcomes and possible influencers in a practical way because it takes into account not only the expected, but also the unexpected. The predictive, statistical modeling within this article helped to determine important reinforcement effects as well as a more generalized treatment effect size that could apply to middle school students as a whole. Essential conclusions can be reached from the data: (a) the number of participants can impact treatment practicality and (b) group structure and variation can influence treatment results. Participants and the treatment variation are major factors in whether or not a treatment has practical significance. Statistically speaking, small sample sizes are not ideal for calculations. The more unique the group characteristics are, the less likely the treatment effects would have the same effects on a generalized sample. The actual group selection and treatment group determination can influence treatment effect size, and enough standardization in research methods should be in place so that statistical results are more likely to align accurately with previous results.

The example models shown in this article help represent concepts beyond words or emotional appeal. A model can be generated which yields results that educators and researchers can use. Determining the appropriate supports based on students' needs is essential to do, and creating better standards that give instructional flexibility is a step in the right direction. The results of this article provide more complexity to the issue of incentive use in schools. It ultimately provides strong encouragement for practical modeling of causes and effects, which is of interest to those who like to utilize rigorous, credible strategies for meeting and exceeding expectations.

Appendix A

Strategy Checklist (Validity and Reliability)

No of Internal Validity Issues Addressed: /10 (10 Items)

- ☐ History
- ☐ Maturation
- ☐ Regression
- ☐ Selection
- ☐ Mortality
- ☐ Diffusion of treatment
- ☐ Compensatory/resentful demoralization
- ☐ Compensatory rivalry
- ☐ Testing
- ☐ Instrumentation

No of External Validity Issues Addressed: /3 (3 Items)

- ☐ Interaction of selection and treatment
- ☐ Interaction of setting and treatment
- ☐ Interaction of history and treatment

No of Suggested Strategies Employed (Validity): /13 (13 Items)

- ☐ Triangulation
- ☐ Member Checking
- ☐ Peer Review
- ☐ Detailed Description
- ☐ Bias Clarification (if warranted)
- ☐ Presentation of discrepancies
- ☐ Time in Field
- ☐ Debriefing

DOI: 10.5281/zenodo.13744090

☐ Auditing

☐ Replication

☐ Theory/Theme Generation

☐ Coding

☐ Other

No of Suggested Strategies Employed (Reliability) /6 (6 Items)

☐ Instrument Consistency (Time)

☐ Instrument Consistency (Type)

☐ Multiple Readings

☐ Statistical Analyses

☐ Agreement Protocols

☐ Other

TOTAL: /32 = ____%

Appendix B

Modeling Methods Checklist

Checklist: Multivariate Mixed-Effects Modeling

1. Gather Data

A. Choose relevant hardware and software for organization.

B. Create a data table of information.

2. Create the Model(s)

A. Create variables based on your data that are compatible with the software.

B. Construct the relevant model or models.

3. Test Assumptions

A. Linearity

• Plot residuals of relevant variables in numeric form.

• Create transformations for non-linear continuous variables.

B. Collinearity/Covariance

• Determine if relationships and covariance exist with any effects.

• Have categorical variables as numerics in correlations.

• Determine if models fit within the parallel, equal, or no model assumptions.

C. Independence/Influential Data

• Determine if groups exist completely independent from each other.

• Explore possible influences using additional commands found in R.

D. Equal Variances

• Aggregate effect sizes and determine appropriate variables for all studies.

• Run heterogeneity tests, determine z-scores, and run transformations if needed.

E. Normality

• Determine univariate normality.

• Determine multivariate normality/residual normality.

4. Test the Model(s)

A. Use multiple rounds (at least 3) for retesting assumptions and models.

B. Perform ANOVAs and/or ANCOVAs for clarification.

5. Interpret the Data

A. Identify corroborative analysis, plots, supportive resources, and fact-checking procedures.

B. Document possible attributions/influences if they exist.

References

1. 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
2. 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.
3. Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*(3), 260-267.
4. Andreas, S. (2012). *Transforming your self: Becoming who you want to be.* Real People Press: Boulder, CO.
5. Averill, O. H., & Rinaldi, C. (2013). *Research Brief: Multi-tier System of Supports (MTSS).* Urban Special Education Leadership Collaborative. https://www.researchgate.net/publication/257943832
6. 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
7. 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. doi: 10.1901/jeab.2010.93-455
8. Bowman, A.W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations* (Google Books Version). Oxford University Press, Oxford. https://books.google.com/books?id=7WBMrZ9umRYC
9. Cangur, S., Sungur, M. A., & Ankarali, H. (2018). The methods used in nonparametric covariance analysis. *Duzce Medical Journal, 20*(1), 1-6.
10. Christensen, R.H.B. (2016). A tutorial on fitting cumulative link models with the ordinal package. https://rdrr.io/rforge/ordinal/f/inst/doc/clm_tutorial.pdf
11. Christensen, R.H.B. (2019). A tutorial on fitting cumulative link mixed models with clmm2 from the ordinal package. https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf
12. Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods Approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.
13. Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

DOI: 10.5281/zenodo.13744090

14. 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)
15. 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
16. 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.
17. 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. https://pdfs.semanticscholar.org/1870/ad57056432dd3ddb787338795 69e213bab13.pdf?_ga=2.150937136.709877635.1569114922-885223096.1569114922
18. 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
19. Dreger, K. C., & Downey, S. (2021). Meaningful consequences: Determining the relevance of instructional reinforcement in education. *The International Journal of Pedagogy and Curriculum, 28*(2), 65-84. https://doi.org/10.18848/2327-7963/CGP/v28i02/65-84
20. Dreger, K. C., & Downey, S. (2022). Reinforcement practicality for middle school students: A meta-analysis [Adobe Reader Version]. *Journal of Inquiry and Action in Education, 11*(1), 1-35. https://digitalcommons.buffalostate.edu/jiae/vol11/iss1/3
21. Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement* [Google Books Version]. http://books.google.com/books
22. Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate Behavioral Research, 51*(2-3), 207-219. https://doi.org/10.1080/00273171
23. Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1), 93–99. https://doi.org/10.1598/RRQ.41.1.4
24. Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202
25. 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)
26. Gordon, K. R. (2019). How mixed-effects modeling can advance our understanding of learning and memory and improve clinical and educational practice. *Journal of Speech, Language & Hearing Research, 62*(3), 507–524. https://doi.org/10.1044/2018_JSLHR-L-ASTM-18-0240
27. Gu, C., & Ma, P. (2005). Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection. *Journal of Computational and Graphical Statistics, 14*(2), https://doi.org/10.1198/106186005X47651
28. 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)
29. 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.
30. Hayden, L. (2018, August 9). Principal Component Analysis in R (Web Log Tutorial). https://www.datacamp.com/community/tutorials/pca-analysis-r
31. 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
32. 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)
33. Howley, A., Allan, D., Howley, N., & Furst, T. (2023). *All Means All…Maybe: MTSS Policy and Practice Across States in the United States.* Minneapolis: University of Minnesota, TIES Center. https://publications.ici.umn.edu/ties/mtss-policy-and-practice/mtss-policy-and-practice
34. Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*(6), 1-24. https://doi.org/10.18637/jss.v075.i06
35. Kuznetsova, A., Brockho, P. B., Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26. https://doi.org/ 10.18637/jss.v082.i13
36. Lawson, A. (1983). Rank Analysis of Covariance: Alternative approaches. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *32*(3), 331-337.
37. 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.
38. 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
39. 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.

DOI: 10.5281/zenodo.13744090

40. Maxwell, J. A. (2012). *Qualitative research design: An interactive approach* (3rd ed.). Thousand Oaks, CA: Sage.
41. 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.
42. 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.
43. McNeish, D., & Kelley, K. (2018, June 4). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000182
44. McSweeney, M., & Porter, A. C. (1971). Small sample properties of nonparametric index of response and rank analysis of covariance (Office of Research Consultation Occasional Paper No. 16). East Lansing: Michigan State University.
45. Mertler, C. A., & Vannatta, R. A. (2013). *Advanced and multivariate statistical methods: Practical application and interpretation* (5th ed.). Glendale, CA: Pyrczak Publishing.
46. 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)
47. 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
48. National Technical Assistance Center on Positive Behavior Interventions and Support. (2017). *Technical guide for alignment of initiatives, programs, practices in school districts.* https://www.pbis.org/resource/technical-guide-for-alignment-of-initiatives-programs-and-practices-in-school-districts
49. Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software, 50*(12), 01-23.
50. Novak, G., & Hammond, J. (1983). Self-reinforcement and descriptive praise in maintaining token economy reading performance. *Journal of Educational Research, 76*(3), 186-189.
51. Olejnik, S., & Algina, J. (1984). Parametric ANCOVA and the Rank Transform ANCOVA when the data are conditionally non-normal and heteroscedastic. *Journal of Educational Statistics,9*(2), 129-149. https://doi.org/10.2307/1164717
52. Olejnik, S. F., & Algina, J. (1985). A review of nonparametric alternatives to analysis of covariance. *Evaluation Review*, *9*(1), 51–83. https://doi.org/10.1177/0193841X8500900104
53. O'Rourke, K. (2007). An historical perspective on meta-analysis: Dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine, 100*(12), 579–582. https://doi.org/10.1177/0141076807100012020
54. Paul, J., & Barari, M. (2022). Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how? *Psychology & Marketing, 39*(6), 1099-1115. https://doi.org/10.1002/mar.21657
55. 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.
56. Preston, A. I., Wood, C. L., & Stecker, P. M. (2015). Response to Intervention: Where it came from and where it's going. *Preventing School Failure, 60*(3), 1-10. https://doi.org/10.1080/1045988X.2015.1065399
57. Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association, 62*(320), 1187 – 1200.
58. School Superintendents Association, Children's Defense Fund. (2014). Positive behavioral supports. https://www.childrensdefense.org/wp-content/uploads/2018/06/positive-behavioral-supports-1.pdf
59. Schweyer, A. (2021). Academic research in action: The psychology of points reward programs. https://theirf.org/research/academic-research-in-action-the-psychology-of-points-reward-programs/3232/
60. Schweyer, A. (2022). Academic research in action: Social reinforcement and peer recognition networks. https://theirf.org/research/academic-research-in-action-social-reinforcement-and-peer-recognition-networks/3280/
61. Scrucca, L. (2001). Nonparametric kernel smoothing methods. The sm library in Xlisp-Stat. *Journal of Statistical Software, 6*(7), 1-49. https://doi.org/10.18637/jss.v006.i07
62. Self-Brown, S. R., & Mathews, I. (2003). Effects of classroom structure on student achievement goal orientation. *Journal of Educational Research*, *97*(2), 106-111.
63. Siegel, A. (2021). Department of Education: Applications for new awards; Teacher
64. and school leader incentive program. *Federal Register, 86*(129), 36262-36269.
65. Simon, S. J., Ayllon, T., & Milan, M. A. (1982). Behavioral compensation. *Behavior Modification*, *6*(3), 407-420.
66. 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. https://dropoutprevention.org/wp-content/uploads/2015/05/Simonsen_Fairbanks_Briesch_Myers_Sugai_2008.pdf
67. Skinner, B. F. (1953). *Science and Human Behavior* [Google Books version]. http://books.google.com/books
68. Strahan, D. B., & Layell, K. (2006). Connecting caring and action through responsive teaching: How one team

DOI: 10.5281/zenodo.13744090

accomplished success in a struggling middle school. *The Clearing House, 79*(3), 147-153.
69. Sugai, G., & Horner, R. H. (1999). Discipline and behavioral support: Preferred processes and practices. *Effective School Practices, 17*(4), 10-22.
70. Sugai, G., & Simonsen, B. (2012). Positive Behavioral Interventions and Supports: History, defining features, and misconceptions. http://pbisaz.org/wp-content/uploads/2013/01/PBIS_History_June19_2012.pdf
71. 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.
72. Swain-Bradway, J., Lindstrom Johnson, S., Bradshaw, C., & McIntosh, K. (November 2017a). What are the economic costs of implementing SWPBIS in comparison to the benefits from reducing suspensions? https://assets-global.website-files.com/5d3725188825e071f1670246/5d76c00cb9339d5f3f267ee7_economiccostsswpbis.pdf
73. Swain-Bradway, J., Putnam, R., Freeman, J., Simonsen, B., George, H. P., Goodman, S., Yanek, K., Lane, K. L., & Sprague, J. (December 2017b). *PBIS technical guide on classroom data: Using data to support implementation of positive classroom behavior support practices and systems.* https://assets-global.website-files.com/5d3725188825e071f1670246/5d9cb4fa139dea541b717452_PCBS%20Data%20Brief%2010.7.19.pdf
74. Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York: Allyn and Bacon.
75. 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)
76. Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology, 52*(1), 124–139. https://doi.org/10.1111/psyp.12299
77. 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.*
78. Tsangari, H. & Akritas, M. G. (2001). Nonparametric ANCOVA with two and three covariates. *Journal of Multivariate Analysis, 88* (2004), 298–319.
79. 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.
80. 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
81. U. S. Department of Education. (2021). Elementary and secondary school emergency relief programs governor's emergency education relief programs. https://oese.ed.gov/files/2021/05/ESSER.GEER_.FAQs_5.26.21_745AM_FINALb0cd6833f6f46e03ba2d97d30aff953260028045f9ef3b18ea602db4b32b1d99.pdf
82. von Ravensberg, H., & Blakely, A.W. (June 2017). *Guidance for states on ESSA state plans: Aligning the school climate indicator and SW-PBIS.* https://assets-global.website-files.com/5d3725188825e071f1670246/5d8a8733506a9e5e3864a113_Guidance%20for%20States%20on%20ESSA%20State%20Plans.pdf
83. Walker, H. M., Horner, R. H., Sugai, G., Bullis, M., Sprague, J. R., Bricker, D., & Kaufman, M. J. (1996). Integrated approaches to preventing antisocial behavior patterns among school-age children and youth. *Journal of Emotional and Behavioral Disorders, 4*(4), 194-209. https://doi.org/10.1177/106342669600400401
84. Whitney, T., Cooper, J. T., & Lingo, A. S. (2018). Using a token economy combined with a mystery motivator for a student with autism exhibiting challenging behavior. *Kentucky Teacher Education Journal: The Journal of the Teacher Education Division of the Kentucky Council for Exceptional Children, 5*(2), Article 1. https://digitalcommons.murraystate.edu/cgi/viewcontent.cgi?article=1013&context=ktej
85. Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications.* http://arxiv.org/pdf/1308.5499.pdf
86. 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.
87. 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)
88. 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)
89. Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R (ebook version). New York, NY: Springer.* https://www.researchgate.net/

DOI: 10.5281/zenodo.13744090