Application of Behavioral Economics to Education

Ji Young Kim

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022
Abstract

Application of Behavioral Economics to Education

Ji Young Kim

Behavioral economics concepts can advance understanding, predicting, and controlling complex human behavior. Although there have been numerous attempts to apply behavioral economics concepts to applied settings, there are a limited number of experiments conducted in educational settings. The purpose of the studies in this paper is to apply well-grounded scientific findings in behavioral economics to education to advance our understanding of teaching and learning and to develop cost-effective interventions. Study 1 (Kim et al., 2021) was a systematic review and meta-analysis on 24 token economy studies conducted in general and special education classrooms from kindergarten to 5th grade between 2000 and 2019. Eight token economy components and effect sizes were identified for each study and compared across different classroom types. The results showed that the token economy intervention yielded large effect sizes for both general and special education classroom types. There were differences in the usage of token components including backup reinforcer types, token production rate, and exchange production rate based on classroom types. In Study 2, the researchers conducted two experiments to develop a rapid, user-friendly assessment of delay discounting for educators interested in quantifying sensitivity to reward delays for school-aged children. The first experiment reanalyzed data collected by Reed and Martens (2011) and found that 1-month delay choices predicted student classroom behavior. The second experiment investigated the utility of the 1-month delay indifference point in predicting saving and spending behavior of second-grade students using token economies with two different token production schedules. Collectively,
results showed that the 1-month delay indifference point predicted classroom behavior and children who discounted less and had greater self-regulation, accrued and saved more tokens. In Study 3, the researchers investigated the effects of a classwide progressive delay training procedure on students’ advantageous choice-making behavior, which is the behavior of selecting a larger, delayed outcome over a smaller, immediately available reinforcer. The intervention used progressive delay, which refers to incrementally increasing the temporal delay for a more delayed, advantageous outcome, and choice-making opportunities to promote second grade students’ choice of a larger reinforcer associated with a delay. The results showed mixed results – the intervention was effective at increasing advantageous choice-making behavior for students performing on and below grade-level for math, but the effects were not shown in students performing above grade-level. Study 4 further examined the educational importance of delayed consequences by determining the relationship between verbal behavior repertoires, delay discounting, and academic achievement. Specifically, the researchers used the degree of Incidental-Bidirectional Naming (Inc-BiN) for verbal behavior, degree of tolerance to delayed consequences using hypothetical binary choices involving money for delay discounting, and standardized math and reading scores for academic achievement. The results showed that participants with greater degree of Inc-BiN had significantly greater tolerance for delayed consequences. However, there was no relation found between academic achievement and delay discounting or Inc-BiN. Together, the four studies in this paper successfully (1) translated basic behavioral economics research findings to educational settings and (2) investigated methods that maximize efficiency and effectiveness of practices and tools used in classrooms.

*Keywords*: behavioral economics, token economy, behavior analysis, education.
# Table of Contents

List of Charts, Graphs, Tables, and Illustrations.................................................................vi

Acknowledgements.............................................................................................................viii

Dedication............................................................................................................................x

Chapter 1: Behavioral Economics and Education..............................................................1

1.1 Behavioral Economics Concepts in ABA.................................................................3

1.1.1 Demand Functions.................................................................................................4

1.1.2 Reinforcer Competition.......................................................................................6

1.1.3 Open and Closed Economics..............................................................................7

1.1.4 Delay Discounting..............................................................................................8

1.2 A Need for Application of Behavioral Economics Concepts to Education.................9

Chapter 2: Systematic Review and Meta-Analysis of Token Economy Practices in K-5 Educational Settings, 2000-2019 (Published in *Behavior Modification*).................................................18

2.1 Method.....................................................................................................................21

2.1.1 Data Search.........................................................................................................21

2.1.2 Inclusion and Exclusion Criteria.......................................................................21

2.1.2.1 Stage 1: Keyword Search..............................................................................22

2.1.2.2 Stage 2: Article Search................................................................................23

2.1.2.3 Stage 3: Citation and Reference Search.......................................................23

2.1.3 Data Search Intercoder Agreement.................................................................24

2.1.4 Data Analysis.....................................................................................................25

2.1.5 Data Analysis Intercoder Agreement...............................................................25

2.1.6 Effect Size Calculations....................................................................................26
3.2.1.1 Participants……………………………………………………………….71
3.2.1.2 Setting and Materials………………………………………………….72
3.2.1.3 Procedure………………………………………………………………73
    3.2.1.3.1 Preference Assessment…………………………………………..73
    3.2.1.3.2 Token Economy…………………………………………………..74
        3.2.1.3.2.1 Fixed Token Production Schedule………………………75
        3.2.1.3.2.2 Variable Token Production Schedule……………………75
    3.2.1.3.3 Behavior Rating Inventory of Executive Functioning (BRIEF2)
        Assessment……………………………………………………………..76
    3.2.1.3.4 Brief Intertemporal Choice Task (BRIC Task) ………………….76
    3.2.1.3.5 Data Analysis……………………………………………………77
        3.2.1.3.5.1 Token Economy Data Analysis…………………………77
        3.2.1.3.5.2 Descriptive Analysis……………………………………77
        3.2.1.3.5.3 Statistical Analysis………………………………………78
    3.2.1.4 Interobserver Agreement………………………………………………78
3.2.3 Results and Discussion……………………………………………………78
    3.2.3.1 Analysis of BRIEF2 and the BRIC Task…………………………78
    3.2.3.2 Mean Peak Point and Mean Balance……………………………..80
3.3 General Discussion……………………………………………………………81
    3.3.1 Implications for Future Research……………………………………83
    3.3.2 Implications for Educators…………………………………………84
3.4 Limitations…………………………………………………………………85
3.5 Conclusion………………………………………………………………88
Chapter 4: A Preliminary Evaluation of an Online Classwide Delay of Gratification Training using Progressive Delay in General Education………………………………………………………………………….105

4.1 Method…………………………………………………………………………111

4.1.1 Participants………………………………………………………………………..111

4.1.2 Setting and Materials………………………………………………………113

4.1.3 Dependent Variable………………………………………………………………………..113

4.1.4 Procedure…………………………………………………………………………115

4.1.4.1 Natural Baseline………………………………………………………...115

4.1.4.2 Choice Baseline………………………………………………………...116

4.1.4.3 Delay of Gratification Training………………………………………...117

4.1.5 Experimental Design………………………………………………………………118

4.1.6 Interobserver Agreement and Treatment Fidelity……………………………119

4.2 Results…………………………………………………………………………119

4.2.1 Individual Participants……………………………………………………………120

4.2.2 Group of Participants……………………………………………………………..121

4.2.2.1 Statistical Analysis……………………………………………………………..123

4.3 Discussion………………………………………………………………………124

Chapter 5: Delayed Consequences in General Education through the Lenses of Verbal Behavior and Delay Discounting………………………………………………………………………….146

5.1 Method…………………………………………………………………………………152

5.1.1 Participants and Setting……………………………………………………………..152

5.1.2 Materials……………………………………………………………………………..153

5.1.3 Measurement………………………………………………………………………..154
5.1.3.1 Incidental Bidirectional Naming (Inc-BiN) ..........................154
  5.1.3.1.1 General Procedure ..............................................154
  5.1.3.1.2 Naming Experience ..............................................155
  5.1.3.1.3 Inc-BiN Probes .....................................................155
5.1.3.2 Delay Discounting Assessment .......................................155
5.1.3.3 i-Ready Diagnostic Assessment .......................................157
  5.1.4 Statistical Analysis ......................................................158
  5.1.5 Interobserver Agreement and Treatment Fidelity .................158
5.2 Results ..............................................................................159
5.3 Discussion .........................................................................160
Chapter 6: General Discussion ..................................................174
  6.1 Summary ........................................................................184
List of Charts, Graphs, Tables, and Illustrations

1. List of Behavioral Economics Terms
2. List of Token Economy Components and Definitions
3. List of Publishing Journals
4. Table of Token Economy Components in General Education Classrooms
5. Table of Token Economy Components in Special Education Classrooms
6. Schematic Overview of the Data Search and Data Analysis Process
7. Effect Size Using Improvement Rate Difference
8. 2009 PRISMA Checklist
9. Table with Results of Reanalysis of Reed & Martens (2011)
10. Table with Correlations among Mean Peak Point, Mean Balance, BRIEF2, and Subjective Value of $100 at 1-Month Delay
11. Individual Change of Mean Peak Point and Mean Balance When Token Production Schedule Changes
12. Scatterplots of Mean Peak Point and Mean Balance between Students with and without Disabilities
13. Scatterplots of Mean Peak Point and Mean Balance between Students who Receive Free/Reduced Lunch and Regular Lunch
14. Scatterplots of Mean Peak Point and Mean Balance between Students who are ELL and Not
15. Scatterplots of Mean Peak Point and Mean Balance for Female and Male Students
16. Example of the BRIC Task Datasheet
17. Table of participant Demographic Information
18. Bar Graphs with Percentage of Larger Later Reinforcer Selection per Participant
19. Link Graphs with Number of Math Problems Solved in the Last Three Days of Pre-
   Intervention Natural Baseline, Intervention, and Post-Intervention Natural Baseline per
   Participant

20. Bar Graphs with Percentage of Larger Later Reinforcer (LLR) Selection

21. Line Graph of Mean Number of Math Facts Completed for Group 1, 2, and 3

22. An Example of Slides Displayed for Choice Baseline Sessions

23. An Example of Slides Displayed during Delay of Gratification Training Phase

24. Table with Nonparametric Correlations (Spearman’s rho) among i-Ready Scores, Inc-Bin,
   and Delay Discounting Measures after exclusion criteria applied

25. Example of the BRIC Task Datasheet
Acknowledgements

It is a genuine pleasure to express my deepest gratitude to everyone on this page. First, I would like to acknowledge and give my most heartfelt thanks to my mentor, advisor, and guide, Dr. Fienup who made this journey possible. I have become a better researcher, teacher, and supervisor because of you, and I am thankful for having to know you since day 1 in ABA1. It has been an honor taking this journey with you and thank you for believing in me and for being an example to follow.

I am also grateful for Dr. Greer who continuously taught me the value of loving to learn and how we should teach that value to our students. Thank you for always pushing me to think outside of the box and inspiring me to make the world a better place. Dr. Jahromi, thank you for being a great role model as an inspiring female researcher and for opening my eyes to different types of research. Dr. Weber, you are such an amazing supervisor and thank you for always being there for us when we needed help the most.

My journey would not have been possible without the help of my mentors, Madeline, Georgette, and Lauren. Thank you for shaping me to become a better teacher and for being someone I can look up to every single day. I am also so thankful for my mentees, Mary-Kate, Josie, Sarah, Amanda, Hayley, Kalie, and Cass, who taught me more than I have ever taught them. I am grateful for the fun times we had in the classroom and working as a team to help students learn and grow.

I would not be here without my Hillcrest family who made work my favorite place to be. Gaby, I know I say this every day, but I truly mean it when I say I could not have made it this far without you at Hillcrest and at Teachers College. Tanya, I am so grateful for having to be able to call an amazing woman and teacher like you my friend. Thank you to both of you for being
genuine friends, inspiring teachers, and simply amazing people. Carol, you are the reason why Hillcrest is my favorite place to be. Thank you for always having open arms to listen to my days and being there as our mama bear.

Words cannot describe how thankful I am for my family – my mom, dad, and my brother, Min Young. You all have always been there for me no matter what and your endless love and support made me the woman I am today. I will save my big words for the day we all meet and have dinner together at home to celebrate.

Lastly, thank you Da Han for being on this journey with me since day 1. You were always there like a rock to wipe my tears, give me hugs, and help me smile when I did not believe in myself. Your jokes and smiles made me forget all the troubles and I thank whoever brought us together. Thank you for being my best friend and for being you.
Dedication

In dedication to my mom, Sung Hee Park, Ph.D. who is a constant reminder of what all women can be and an inexhaustible source of inspiration.
Chapter 1: Behavioral Economics and Education

In recent years, there has been an increase in students receiving special education services within the education system. From 2009-10 through 2019-20 school years, the number of students ages between 3 to 21 years receiving special education services under the Individuals with Disabilities Education Act (IDEA) increased from 6.5 million (13% of total public school enrollment) to 7.3 million (14%; United States Department of Education, 2021). Among these students, the disability types include specific learning disability (33% of students who received special education services), speech or language impairment (19%), and other health impairments (15%). Each disability type provides a general profile of the student and helps educators modify their teaching based on students’ specific needs. However, it is often difficult to categorize a student into one type of disability and the “types” of disabilities assume that disabilities are relatively fixed and stable. The general descriptions or categories of disabilities overlook individual complexities and may lead to educators attributing behavioral problems to a particular type of disability (e.g., a teacher attributing students’ slow rate of learning to his/her classification of specific learning disability) rather than analyzing the contingencies of the problem behavior and utilizing research-based instructional strategies to address the target behavior. As behavior scientists, it is important to analyze and quantify the target behavior in an objective manner to predict, control, and understand the target behavior to shape it in a more desirable direction. As educators, it is important to apply the scientific findings and utilize evidence-based teaching technologies as a strategic scientist (Greer, 1991) to adequately address the individual student needs.

Applied behavior analysis (ABA) has a strong foundation of high-quality research to systematically change socially significant behavior (National Autism Center, 2015). Since
Skinner’s application of technology to teaching (Skinner, 1968), the education system started to gradually incorporate scientific teaching strategies and behavioral interventions into everyday practice. Behavior scientists have used rigorous analytic research methods to support the effectiveness of these interventions. Within ABA, specifically, it is important to note that many applied practices stem from basic research where experiments are conducted in a tightly controlled setting such as laboratories with nonhuman animals or human subjects. Behavior scientists continuously ask how the basic findings can be translated into practice and better inform clinicians to address client needs and educators to advance teaching and learning (Cooper et al., 2019; Greer, 2002).

Translational research typically begins with findings from controlled laboratory studies which are later replicated to clinical populations and settings (Lerman, 2003). The findings are evaluated for development of effective technologies. Systematic line of translational study involves (a) bridging basic and applied outcomes that are germane to clinical practice, (b) evaluating variables that promote transfer of laboratory-based technologies into applied settings, and (c) demonstrating the application of these technologies by practitioners, parents, and teacher as part of routine service (Lerman, 2003). Translational research is important in the field of ABA because it is directly in line with at least two of the seven dimensions of ABA, Applied and Analytic – behavior scientists analyze behavior in a controlled manner across applied settings where “immediately important” (p. 93) behavior and/or stimuli are being studied (Baer et al., 1986, 1987). According to Greer (2002), “The principles of the basic science of the behavior of the individual and tactics from the applied research are used to teach educationally and socially significant repertoires” (p.14). The application of the basic research findings extends the scope of the theoretical underpinnings of behavior science and enrich the technologies used in applied
settings. Critchfield (2011) argued that we need coordinated collaborations between basic and applied researchers because it is “genuinely useful to do so” (Vollmer, 2011, p.31). Thus, it is paramount to bridge the gap between basic science and practice to develop well-informed techniques grounded in science.

The beginning of this paper discussed the importance of using science to guide our practices in educational settings to highlight that there are still areas within basic experimental research that can advance our understanding of classroom contingencies and responses. Behavioral economics is an emerging subfield of Psychology, Economics, and Behavior Analysis. Response-reinforcer relations are often investigated within behavioral economics, but the topic has largely remained in the basic realm, with implications for educational practices, although few direct applications. In an educational setting, response and reinforcement are constantly in play and manipulated by educators to advance learning. For example, teachers often give out small prizes or praise to reward a student’s achievement. Researchers developed scientific techniques such as token economy (Kazdin & Bootzin, 1972) to manipulate the response-reinforcer arrangements that affect student performance or learning (Ivy et al., 2017; Kim et al., in press). Despite a plethora of research conducted in classrooms and in laboratories, there is a need for translational research applying behavioral economics concepts to education given the clear implications in applied settings. This line of research will better inform educators on classroom contingencies and on the development of effective and efficient interventions to improve meaningful behaviors such as decision making or allocation of behavior to advance teaching and learning (Reed et al., 2011). The next section will discuss the four behavioral economics concepts

1.1 Behavioral Economics Concepts in ABA
The field of behavioral economics originally rooted from a group of researchers’ attempts at understanding and modeling irrational consumer choices (Reed et al., 2013). While there has been mentalistic or psychological attempts of explaining irrationality, operant principles help explain irrationality and behavioral economics concepts (Madden, 2000; Reed et al., 2013; Skinner, 1953). This makes behavioral economics inherently conceptually systematic, which refers to when interventions and technologies are consistent with the operant principles of ABA (Baer et al., 1968). Recently, there has been increased attention to behavioral economics within the field of behavior analysis (Bickel et al., 1995). Despite the extensive basic experimental studies conducted and greater interest of the field, the application of behavioral economics to educational or therapeutic settings remains limited. Nevertheless, behavioral economics is “a ubiquitous concept rather than a behavior change procedure” (Reed et al., 2013, p.51) given that the principles of behavioral economics are often intentionally or unintentionally incorporated into interventions implemented within applied settings.

For decades, behavior scientists have suggested that behavioral economics concepts can advance understanding, predicting, and controlling complex human behavior. In response, Hursh (1980, 1984) identified three behavioral economics concepts commonly used behavior science: (b) demand function, (c) reinforcer competition, and (d) open and closed economies. Later, Mazur (1987) added the concept of delay discounting. Researchers have attempted to apply these basic concepts to applied settings. In this section, text describes each concept and its applied use.

**1.1.1 Demand Functions**

The first major behavioral economics concept is demand functions, which is the degree to which the level of consumption remains stable across unit price increases (Hursh, 1980, 1984). In behavior science, consumption refers to the number of reinforcers that can be obtained within
a session, and unit price refers to the number of responses required to obtain a reinforcer (see Table 1 for definitions of behavioral economics terms). There are two major types of demand functions – elastic and inelastic. Demand functions are inelastic when the change in unit price does not affect the level of consumption. Demand functions become elastic when the unit price becomes too high and thus the level of consumption decreases. For example, a student’s demand function is inelastic when she is willing to solve 1, 2, 5, and 10 math problems per reinforcer and earn 10 tokens in each session. However, when the number of problems required per reinforcer becomes higher than 10, the student will begin to engage less in the target response and earn fewer tokens within a given session. In this example, the student’s behavior becomes elastic at a unit price higher than 10.

A seminal study by Hursh et al. (1988) demonstrated how demand curves plot for preferred commodities. Hursh and colleagues manipulated the number of responses needed to obtain access to food for rats working for food pellets. The rats’ consumption of food and numbers of responses initially increased but reached the point of elasticity where consumption and responses decreased. Researchers have recently begun using the concept of demand functions to applied settings. These studies examined reinforcer efficacy to inform practitioners on efficient use of reinforcers to maintain responding and reduce the cost of reinforcers. For example, Borrero et al. (2007) conducted a unit-price analysis across six individuals with severe problem behavior and found that these participants’ behaviors were elastic when plotted on a demand function. This means that when the reinforcer price increased, the participants engaged less in the target appropriate behavior. Another study by Roane et al. (2001) evaluated the reinforcing efficacy of different stimuli as reinforcers under a progressive-ratio (PR) schedule of reinforcement and found that there is differential effectiveness of the stimuli as reinforcers as
response requirements increased. Studies investigating the utility of PR schedules of reinforcement on responding in an applied setting have been extended further since then (DeLeon et al., 2000; Kodak et al., 2007; Roane et al., 2005). The studies collectively inform us that the application of demand functions help us understand how response and reinforcement interplay in an applied setting, and that there is a need for more research on direct application of this concept to education.

1.1.2 Reinforcer Competition

The second behavioral economics concept is reinforcer competition. In the real world, there are often more than one reinforcer at the same time and these reinforcers interact with one another. This means that different reinforcers in the environment affects human behavior simultaneously. For example, when adults work, promotion and monetary reinforcement both function to increase the level of responding at work (i.e., work productivity). Understanding the relationship among these reinforcers is important because stimuli often become part of complex human behavior and new reinforcers or stimuli which expand one’s environment in turn select behaviors (Greer, 2020). Reinforcer competition stemmed from this idea and is referred to as how reinforcers interact with one another in terms of their consumption and unit price (Hursh, 1980, 1984). Three types of relationship between reinforcers exist – substitutable, complementary, and independent. Reinforcers are substitutable when the increase in the unit price of one reinforcer increases the consumption of another reinforcer (e.g., smiley stickers and star stickers). Reinforcers are complementary when the consumption of two reinforcers increase and decrease together regardless of the price change (e.g., pencils and erasers). Reinforcers can also be independent, which means that they do not interact with one another despite the changes in unit price (e.g., pencils and shoes).
One study looked at substitutability in rats by providing a choice between two liquids where both were associated with an equal and low response requirement (Rachlin et al., 1976). When the price was equal, the rats preferred liquid 1, but as the liquid 1’s price increased and liquid 2 remained relatively low, the rats started to prefer liquid 2. Therefore, liquids 1 and 2 were substitutes. Researchers have also studied the dynamics of reinforcers within applied settings. Salvy et al, (2009) explored how food and social time with peers interact with one another and showed that food and social time with unfamiliar were substitutable, meaning that if the price of food increased, students began to engage in behaviors associated with earning social time with unfamiliar peers. They also found that social time with familiar peers was always more reinforcing than food. Understanding reinforcers as substitutable and complementary also have been suggested to better inform us on the development of obesity treatment programs (Carr & Epstein, 2020), extending the utility and applicability of reinforcer competition to fields outside of behavior analysis.

1.1.3 Open and Closed Economies

The third behavioral economics concept is open and closed economies. Value of reinforcers change based on the relative availability of it within or outside of the target system (Hursh, 1980, 1984). Depending on the reinforcer availability within or outside the system, rate of responding is differentially affected. An open economy refers to when reinforcers are available inside and outside the target system. On the other hand, a closed economy refers to when reinforcers are available only inside the target system. For example, in a closed economy, a student can use tokens to exchange for iPad time only at school (target system) contingent on completing assignments. In an open economy, this student can exchange for iPad time at school and at home.
In the past, many researchers studied the principle of open and closed economies using nonhuman animal subjects. For example, LaFiette and Fantino (1989) used pigeons to compare the effectiveness of the two different types of economies on the rate of responding (i.e., lever pressing). In the open economy condition, pigeons were provided 80% of free body-weight food the entire time while in the closed economy condition, pigeons were given access to food after 23.5 hours of no food. The rate of responding was higher in a closed economy compared to an open economy. The findings were extended to applied settings where researchers showed that students emitted more responses under the closed economy system when PR was used for reinforcer delivery (Roane et al., 2005). However, this study was conducted in an unused classroom, suggesting the need for studies conducted in general education settings.

1.1.4 Delay Discounting

The last behavioral economics concept is delay discounting which stems from the idea that humans are often myopic when faced with delayed consequences (Reed et al., 2013). Delay discounting is defined as systematic devaluation of reinforcement as a function of increasing temporal delays. When individuals are faced with a more advantageous reinforcer than can be obtained later (Larger Later Reinforcer; LLR) and a less advantageous reinforcer that can be obtained sooner (Smaller Sooner Reinforcer; SSR), they struggle to continue to choose the LLR. At some point, individuals switch their choices from LLR to SSR as the delay increases. Altogether, delay discounting refers to the behavioral pattern where the value of the reinforcer decreases as a function of increase in delay.

Mazur (1987) first formally tested and plotted delay discounting using pigeons. He introduced two concurrent levers where one was associated with LLR and another with SSR. He then progressively increased the delay to LLR when the pigeon selected LLR. When the pigeon’s
responses were plotted, there was a hyperbolic function between the delay and reinforcer value. Researchers started to extend the utility of delay discounting to humans and developed delay discounting assessments. The most widely used assessment was developed by Rachlin et al. (1991) where individuals are presented with two hypothetical monetary choices (e.g., $1,000 in a week vs. $10 right now). As individuals allocate their selection responses between the two concurrent choices, the choices are plotted into a hyperbolic function. Then, researchers derive numbers such as area under the curve (AUC) and the k-value (i.e., the discounting rate) to quantify how fast individuals discount the reinforcer value given a delay. One study applied the principle of delay discounting to an educational setting. Reed and Martens (2011) compared sixth graders’ rate of discounting to their on-task behavior across delayed and immediate reward conditions and found that students with lower discounting rates (i.e., more tolerant to delays) engage in more on-task behavior under the delayed reward condition. A recent meta-analysis revealed that this was the only study conducted in a classroom setting applying delay discounting (Staubitz et al., 2018), which highlights a further need for application of delay discounting within educational settings.

1.2 A Need for Application of Behavioral Economics Concepts to Education

Behavioral economics affects a wide range of important applied domains within and outside of behavior analysis. Although there have been numerous attempts to apply behavioral economics concepts to applied settings, there are a limited number of experiments conducted in educational settings (Reed et al., 2013; Staubitz et al., 2018). Every day, teachers should apply existing science and technology of teaching as a strategic scientist (Greer, 1991). As an educator and behavior scientist, it is important to approach problems scientifically, efficiently, and effectively. Educational settings often involve limited resources such as time, funds, supplies,
and even human labor (e.g., teachers and assistants), which makes “cost-effective empirically supported interventions” (Reed et al., 2013, p.37) vital. Behavioral economics has much to offer especially given its strong empirical underpinnings in behavior science for effective change and efficient manipulation of environmental contingencies to promote positive behavior change.

The series of studies in this paper is a journey of an educator and a behavior analyst applying well-grounded scientific findings in behavioral economics to education to advance our understanding of teaching and learning and to develop cost-effective interventions. The aim of the four studies in this paper is two fold – (1) to translate basic behavioral economics research findings to educational settings and (2) to conduct studies that maximize efficiency and effectiveness of practices and tools used in classrooms. The first study was a systematic review and meta-analysis of token economy arrangements across general and special education classrooms from kindergarten to 5th grade. I aimed to compare different token economy components to identify an optimal arrangement of token economy depending on the classroom type. That way, educators who use token economies can more effectively tailor the system to their classroom needs and maximize effectiveness of the system. The second study introduced a brief, user-friendly measure for educators interested in quantifying sensitivity to reward delays for school-aged children. An efficient assessment of delay discounting would help educators quickly determine individual students’ level of sensitivity to delay, which would help them arrange the classroom contingencies. The third study evaluated the effects of an online, group-based delay of gratification training to teach students to engage in more math problems to obtain a larger reinforcer. I used an intervention procedure that could increase the number of participants contacting the positive behavior change within the same time frame. The intervention addressed students’ academic and self-management repertoires at the same time,
highlighting the efficiency of the intervention. In the fourth experiment, I analyzed the relation between academic achievement, delay discounting, and verbal behavior to determine the implications of delayed consequences in general education settings. With limited line of research evaluating the correlating variables with implications on delayed consequences, this study represents a first step in bridging the gap between the two distinct realms within behavior analysis – verbal behavior and delay discounting. Specifically, I examined the standardized math and reading scores for academic achievement, degree of delay discounting using hypothetical binary questions involving monetary choices, and degree of Incidental Bidirectional Naming (Inc-BiN) for verbal behavior.
References


Table 1

Definitions of behavioral economics terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumption</td>
<td>The amount of a commodity obtained in a session of observation.</td>
</tr>
<tr>
<td>Unit price</td>
<td>A ratio of costs and benefits (i.e., response requirements/reinforcer).</td>
</tr>
<tr>
<td>Cost</td>
<td>Requirement an individual must meet to obtain a commodity.</td>
</tr>
<tr>
<td>Benefit</td>
<td>The amount of a commodity that can be earned.</td>
</tr>
<tr>
<td>Commodity</td>
<td>Reinforcer an individual will work to earn.</td>
</tr>
</tbody>
</table>

Note. The definitions were adapted from Hursh et al. (2013) and Reed et al. (2013).

(Published in *Behavior Modification*)

Outline

Token economy systems have been widely used as an evidence-based classroom management strategy to reinforce and improve prosocial responses. While token economies have been widely applied to educational settings, there have been mixed results regarding the effectiveness depending on the classroom type. To better understand the components contributing to the effectiveness, the researchers analyzed 24 token economy studies conducted in general and special education classrooms from kindergarten to 5th grade between 2000 and 2019. Eight token economy components and effect sizes were identified for each study and compared across different classroom types. The results showed that the token economy intervention yielded large effect sizes for both general and special education classroom types. There were differences in the usage of token components including backup reinforcer types, token production rate, and exchange production rate based on classroom types. Implications for future research and practice for educators and clinicians are discussed.

*Keywords:* behavioral economics, general education classroom, special education classroom, token economy.

Educators are encouraged to use research-based strategies to manage student behavior in school settings (Gresham, 2004). Token economies have been identified as an effective evidence-based classroom management strategy (Simonsen et al., 2008) and a well-established psychological procedure by the American Psychological Association, Task Force on Promotion and Dissemination of Psychological Procedures (1993). In classrooms, educators implement token economies by choosing specific behaviors to increase, delivering tokens contingent on students emitting that behavior, and students exchanging tokens for various goods, activities, or privileges at a scheduled time (Ayllon & Azrin, 1968). A basic token economy consists of six primary features or components (Hackenberg, 2009; Ivy et al., 2017; Kazdin & Bootzin, 1972) including (a) target behavior, (b) tokens that function as conditioned reinforcers, (c) backup reinforcers, (d) token-production schedule, (e) exchange-production schedule, and (f) token-exchange schedule. These components are essential in making a token economy system flexible and adaptable to various settings and populations.

Given this flexibility, several studies have demonstrated successful application of token economies across different settings and populations with varying age, gender, disabilities, and disorders (Doll et al., 2013; Ivy et al., 2017; Kazdin & Bootzin, 1972; Soares et al., 2016). Token economies have been implemented in primary and secondary schools (Maglio & McLaughlin, 1981; McLaughlin & Malaby, 1977), universities and colleges (Boniecki & Moore, 2003), prisons (Kazdin, 1977), detention centers (Bippes et al., 1986), inpatient programs (Milby, 1975), and residential treatment center (Murray & Sefchik, 1992). Further, there have been systematic reviews and meta-analyses on token economy documenting the effectiveness of this
technology. Previous reviews have focused on token economy use with specific populations, including individuals with schizophrenia (Dickerson et al., 2005), students with challenging behavior (Maggin et al., 2011), and psychiatric inpatients (Milby, 1975). Reviews also highlighted the historical and modern use of token economy with suggestions for future development. Topics include issues and barriers with the application of token economy (Kazdin, 1982; Kazdin & Bootzin, 1972; O’Leary & Drabman, 1971), historical development and modern usage (Carlson et al., 1972; Doll et al., 2013; Hackenberg, 2009; Matson, & Boisjoli, 2008), procedural components (Ivy et al., 2017), and effect sizes in classrooms (Soares et al., 2016).

Specifically, Ivy et al. (2017) evaluated token economy studies published between 2000 and 2015 to evaluate the procedural descriptions of token economies implemented in applied settings. The results showed that token economy components were often omitted or vaguely described, which may negatively affect future research or applications. Given that token economies are often used in educational settings, it is imperative to evaluate the procedural components of token economies implemented in classrooms. Further, to date, research suggested mixed results regarding the effectiveness of interventions such as token economies across different educational settings (DuPaul & Eckert, 1997; Soares et al., 2016). While interventions in general had greater effects when implemented in special education classrooms compared to general education classrooms for within-subject design studies (DuPaul & Eckert, 1997), token economies have been shown to have comparable effects in general and special education classrooms (Soares et al., 2016). The findings together suggest that an evaluation of the procedural description and effectiveness of the intervention across general and special education classrooms is needed. Thus, the purpose of this study was to compare token economy
components and the effectiveness of token economies implemented in general and special education classrooms.

In our systematic review and meta-analysis, we calculated effect sizes per study and aggregated effect sizes according to type of classroom (general and special education classrooms). The eight token economy components selected (see Table 1) were an extension of the components identified by Ivy et al. (2017). We incorporated target behavior, token production schedule, exchange production schedule, and token exchange schedule from Ivy et al. (2017) and added types of tokens, types of backup reinforcers, response cost, and unit of contingency to the components to be analyzed. This review extended Ivy et al. (2017) by (1) examining a larger range of token economy components, (2) focusing on token economy studies conducted in K-5 educational settings, (3) focusing on token economy studies published in years between 2000 and 2019, and (4) reporting effect sizes of treatment outcomes. This review focused on K-5 grade level because token economies were most commonly used in elementary school classroom settings (Doll et al., 2013; Ruesch & McLaughlin, 1981). This review also used studies published between 2000 and 2019 to better reflect the updated literature on token economy. This study aimed to address the following two research questions. First, what are the variations of the token economy components that produce better outcomes in each classroom type? Second, are token economy systems effective in each classroom type?

2.1 Method

2.1.1 Data Search

The search included three stages. Figure 1 summarizes the stages with number of articles identified and included in each stage.

2.1.2 Inclusion and Exclusion Criteria
The researchers evaluated the articles identified through the data search procedure to evaluate for inclusion in the analysis. The researchers reviewed the title and abstract first and then the full article to evaluate the inclusion eligibility. To be included, a study had to (a) use a token economy system as the intervention or in a treatment package, (b) implement the intervention in a general or special education setting with students’ grade-level ranging from kindergarten to 5th grade with the age ranging from 3 to 12 years old, (c) be published between the years of 2000 and 2019, and (d) have the intervention conducted in the participants’ regular classroom setting (i.e., not a laboratory setting).

The researchers excluded studies that (a) were literature reviews, conceptual papers, and basic research, (b) implemented a reinforcement system similar to token economies for parent or staff training or drug use treatment, (c) implemented token economies outside of a regular school or classroom setting (e.g., research laboratory setting, private clinic, hospital, detention centers, summer programs, universities, rehabilitation centers, or prison), (d) had the student removed from his or her regular classroom to implement the intervention (e.g., small room, intervention room, alcove), (e) included participants who were adolescents, preschoolers, adults, older population, and non-human animals, and (f) did not specify the type of classroom the intervention was conducted.

2.1.2.1 Stage 1: Keyword Search

The researchers searched PsycINFO and Education Resources Information Center (ERIC) database using the keywords token economy and token reinforcement. No other search terms were used. The researchers set the parameters of the search to identify articles published from January 2000 through December 2019 in peer-reviewed English language journals. The search yielded 819 articles in total (563 for token economy and 256 for token reinforcement).
2.1.2.2 Stage 2: Article Search

Stage 2 consisted of two levels of analysis. The researchers applied the inclusion and exclusion criteria for articles identified in Stage 1 through a two-level analysis. Duplicates were excluded. During the first level of analysis, the researchers reviewed the title and abstract to determine if the article would fit the inclusion criteria. During the second level of analysis, the researchers reviewed the articles identified through the first level of analysis and eliminated those that did not fit the inclusion criteria. The researchers identified 176 articles through the first level of analysis and 20 through the second level of analysis. Thus, there were 20 articles at the end of Stage 2.

2.1.2.3 Stage 3: Citation and Reference Search

The researchers used the 20 articles identified through Stage 2 and conducted a citation search using the Google Scholar database and a hand-reference search by inspecting the reference section of each article to locate additional articles that may have been missed through the first round of search. The two-level analysis was applied for the citation and reference searches. Figure 1 shows the number of articles included in each level of analysis. Articles the researchers were unable to obtain through the database and university interlibrary loan service were excluded from the study. The citation search yielded 15 novel articles and reference search yielded four novel articles. In total, the 39 articles were identified. There were three articles excluded from our review because the researchers were unable to obtain them through the database and university interlibrary loan service, and the authors’ email addresses were unavailable.

Out of the 39 articles, the researchers applied the What Works Clearinghouse (WWC) standards for single-case design (Kratochwill et al., 2010) and identified 24 studies that meet the
evidence standards or meet evidence standards with reservations. This process was used to ensure that the current review only included studies meeting the minimal standards for high quality research. Each study had to meet the evidence standards or meet evidence standards with reservations across standards including (1) systematic manipulation of independent variables, (2) sufficient interobserver agreement (IOA) across phases (i.e., an independent second observer assessing at least 20% of the data points in each condition with a minimal 80% agreement for inter-assessor agreement and 60% for Cohen’s kappa), (3) at least three demonstration of intervention effects at three different points in time, and (4) each phase with at least three data points or at least five repetitions of alternations if an alternating treatments design was used. If a study did not meet all four standards with at least meet evidence standards with reservations status, the study was excluded from the review. Fifteen studies were excluded during this process. The most common reasons of exclusions were in the order of insufficient IOA for each condition, less than three data points in each phase, and demonstration of an intervention effect in less than three points in time. None of the included 24 articles contained multiple experiments. Thus, the researcher analyzed 24 unique studies for the current review.

2.1.3 Data Search Intercoder Agreement

Intercoder agreement (ICA) data were collected for each stage. For Stage 1, a second reviewer independently conducted a keyword search by repeating the procedure described in Stage 1. To calculate ICA, the researchers conducted a total count ICA (Cooper et al., 2019) by dividing the smaller number of found articles by the larger number of found articles and multiplied the result by 100. ICA was 96% for the keyword search. For Stage 2, an independent reviewer repeated the procedure described in Stage 2 by applying the inclusion and exclusion criteria for 60% of articles identified through the first level of analysis. The first author randomly
selected the articles, using a list randomizer website (random.org/lists/). To calculate ICA, the researchers used a trial-by-trial (study-by-study) method, where the number of articles with agreement were divided by the number of items with agreement and disagreements and multiplied by 100. ICA was 94% for Stage 2. For Stage 3, an independent reviewer repeated the procedure described in Stage 3 by conducting the search and applying the inclusion and exclusion criteria for 30% of articles identified through the Citation and Reference Search. To calculate ICA, the researchers used the methods described for Stage 1 and 2. The ICA for the search was 100% and applying the inclusion criteria was 92%.

2.1.4 Data Analysis

Following the identification of 24 studies that meet the evidence standards or meet evidence standards with reservations (Kratochwill et al., 2010), the researchers reviewed each article to extract data on each article that met criteria. Along with the basic study information (i.e., authors, name of the article, year of publication, journal, year), the researchers coded the data on the following categories: (a) types of tokens, (b) types of backup reinforcers, (c) target behavior, (d) response cost, (e) token production rate, (f) exchange-production rate, (g) token-exchange rate, and (h) unit of contingency (see Table 1 for definitions of each component).

2.1.5 Data Analysis Intercoder Agreement

A second reviewer independently analyzed 38% of the articles identified by the first author through the data search. The first author randomly selected the articles, using a list randomizer website (random.org/lists/). Trial-by-trial analyses was used to calculate ICA (Cooper et al., 2019) for each of the following variables: (a) setting, (b) target behavior, (c) token type, (d) backup reinforcer type, (e) response cost, (f) token production rate, (g) exchange production rate, (h) token exchange rate, and (i) unit of contingency. Overall ICA was 95%.
2.1.6 Effect Size Calculations

The researchers calculated the effect size using the improvement rate difference (IRD; Parker et al., 2009) to quantify the strength of the treatment in each study. IRD has several advantages including available confidence intervals, successful applications in hundreds of evidence-based medical research studies, and strong correlation with well-known parametric and nonparametric effect sizes including $R^2$, Kruskal-Wallis $W$ effect size, percent of nonoverlapping data (PND), phi, and percent of all over nonoverlapping data (PAND; Parker et al., 2009). IRD is commonly used in medical research for risk analysis and represents a nonparametric effect size for single-case research based on the risk analysis of baseline and intervention data (Parker et al., 2009). Specifically, the difference between two proportions of data overlap provides IRD. The first proportion represents the number of baseline data overlapping with treatment data, and the second proportion represents the total number of baseline data overlapping with baseline data.

To calculate omnibus IRD for each study, the researchers followed the procedures described by Parker et al. (2009). The researchers excluded secondary measures not targeted through the token economy for IRD calculation (e.g., teacher praise). An IRD value is reported on a scale from 0 to 1.00, and a value less than 0.50 indicated a small effect, a value between 0.51 and 0.70 indicated a moderate effect, and a value greater than 0.71 indicated a large effect (Parker et al., 2009; Rakap, 2015). IRD effect sizes and confidence intervals were hand-calculated first and then cross-checked using an online confidence interval calculator (VassarStats; http://www.vassarstats.net/prop2_ind.html).

2.1.7 Effect Size Intercoder Agreement

ICA data for effect sizes were collected for 25% of the total included articles. The first author randomly selected the articles, using a list randomizer website (random.org/lists/). To
calculate ICA, the researchers used point-by-point method (Cooper et al., 2019). ICA was calculated by dividing the lower effect size by the larger effect size and multiplying by 100. Mean agreement was 99% (range, 97% to 100%).

2.2 Results

2.2.1 General Characteristics of the Literature

The 24 studies included in this review were published in 19 journals during the 2000 to 2019 time period. The journals with the highest number of studies (N = 4; 17%; see Table 2) was *Journal of Behavioral Education*, followed by *Psychology in the Schools*, and *Journal of Applied Behavior Analysis* (N = 3; 13%; see Table 2). Studies were conducted most frequently in general education classroom setting (N = 15; 63%), followed by special education (N = 9; 38%) classroom. Twenty-two out of 24 studies (92%) targeted classroom behavior while 2 studies (8%) targeted academic skills such as math word problems (Alter, 2012) and reading (Shepley et al., 2016).

2.2.2 Question (1): What are the variations of the token economy components that produce better outcomes in each classroom type?

2.2.2.1 General Education Classroom

The search identified 15 studies with a token economy system implemented in a general education classroom setting. Table 3 displays the token economy components of each study conducted in general education classrooms. Overall, points and small marks or items (e.g., tick marks, chips, smiley faces, dots) were implemented as tokens. For backup reinforcers, studies employed edibles and small tangibles most often (N = 7; 47%) followed by mystery rewards/motivators (N = 3; 20%) and activities (N = 2; 13%). Most studies reported that these backup reinforcers were communal rather than individualized. In other words, a group of
students shared the same backup reinforcers instead of having them individualized per student based on individual preferences. Four studies (27%) did not report the specific back up reinforcers (Battaglia et al., 2015; Christensen et al., 2004; Christensen et al., 2007; Petursdottir & Ragnarsdottir, 2019), but three out of the four studies stated that they determined the reinforcers through a functional behavior assessment (FBA), reinforcer survey, or preference assessment. All 15 studies investigated the effects of the intervention on classroom behavior and academically related behaviors – such as following classroom rules, socially appropriate behavior, disruptive behavior, and on- and off-task behavior – rather than academic responding, specifically. Five out of 15 studies (33%) used response cost as part of the independent variable. Three studies that used response cost were comparison studies that used response cost in one of the conditions (Donaldson et al., 2014; Tanol et al., 2010; Wright & McCurdy, 2011).

Regarding the token production rate, the studies mostly targeted instances of disruptive behavior and following rules for a pre-determined length of interval which ranged from 30 s to 50 min. Most studies implemented a 10 min interval (N = 4; 27%) for engaging in the appropriate behavior. One study used specific times for check-ins (Akin-Little & Little, 2004). The exchange production rate showed that students mostly traded-in at the end of an intervention session (N = 8; 53%) or at the end of the day (N = 5; 33%). Some studies provided the number tokens required before exchanging (De Martini-Scully et al., 2000) and a specific time after the intervention (Pennington & McComas, 2017). The token exchange rate was most commonly reported in the number of tokens (N = 9; 60%). Two studies (13%) used a system that allowed the students to trade in if they earned a number of tokens below a set criterion. The system

---

1 The number of studies for backup reinforcers does not match the total number of studies because one study used tangibles and activities as mystery rewards (Robichaux & Gresham, 2014).
incorporated positive punishment operations where teachers gave students a token when the student displayed undesirable behavior (Robichaux & Gresham, 2014; Wright & McCurdy, 2011). Two other studies (13%) allowed the students to trade in if they earned more points than the opponent team or the teacher (Lastrapes et al., 2018; Pennington & McComas, 2017). Two studies did not specify the token exchange rate (Christensen et al., 2004; Christensen et al., 2007). The most commonly used cost of reinforcer was 3 or 5 tokens ($N = 4; 27\%$ each). Nine studies out of 19 ($47\%$) utilized group contingency to provide access to a backup reinforcer.

### 2.2.2.2 Special Education Classroom

The search identified 9 studies with a token economy system implemented in a special education classroom setting. Table 4 displays the token economy components of each study conducted in special education classrooms. All studies implemented points and small marks or items as tokens (e.g., stickers, pennies, tickets). Most studies reported that backup reinforcers were individualized rather than communal. In other words, backup reinforcers were individualized per student based on individual preferences. Edibles and small tangibles were employed the most ($N = 6; 67\%$) as backup reinforcers, followed by activities ($N = 4; 44\%$) and mystery rewards/motivators ($N = 2; 22\%$). Two studies (22%) did not report the type of backup reinforcers (Gann et al., 2015; Shepley et al., 2016), but one of the studies stated that they determined the reinforcers through student records and teacher recommendations (Shepley et al., 2016). Seven of 9 studies (78%) targeted classroom behavior and academically related behaviors while two studies (22%) targeted academic responding such as math word problems.

---

2 The number of studies for backup reinforcers does not match the total number of studies because five studies used multiple types of backup reinforcers (Alter, 2012; Groves & Austin, 2017; McDonald et al., 2014; Wadsworth et al., 2015).
(Alter, 2012) and reading words (Shepley et al., 2016). One out of 9 studies (11%) implemented response cost as part of the independent variable (Musser et al., 2001).

For the token production rate, most studies targeted instances of correct target behavior or engaging in the appropriate behavior for an entire interval which ranged from 20 s to 30 min. Most commonly employed rate was one accurate target response for one token ($N = 4; 44\%$).

The exchange production rate showed that students mostly traded-in when they met a predetermined number of tokens ($N = 4; 44\%$) or at the end of a session, day, or week ($N = 3; 33\%$). Two studies (22\%) did not specify the exchange production rate (Gann et al., 2015; McDonald et al., 2014). The token exchange rate was most commonly reported in the number tokens ($N = 6; 67\%$). The number of tokens ranged from 1 to 25. One study did not specify the token exchange rate (Gann et al., 2015). One out of 9 studies (11\%) implemented a group contingency to provide access to a backup reinforcer (Groves & Austin, 2017).

2.2.3 Question (2): Are token economy systems effective in each classroom type?

Figure 2 displays the effect sizes and confidence intervals of each study in the order listed in Table 3 and 4 with an average effect size for general and special education classroom type. On average, token economy systems had large effect sizes for both classroom types, indicating that token economy interventions were highly effective in decreasing inappropriate (e.g., disruptive behavior) and increasing appropriate (e.g., task engagement) behavior across general education and special education classrooms. Token economy studies conducted in special education classrooms demonstrated a slightly larger average effect size ($M = 0.87$) compared to general education classrooms ($M = 0.83$). Although the average effect size was larger in special education classrooms, average IRD effect sizes in both settings were large.
Additionally, studies conducted in general education classrooms had a slightly higher proportion of large effect size (93%; 14 out of 15 studies) compared to those in special education classrooms (89%; 8 out of 9 studies). The average effect size of general education classrooms may be lower due to one study that showed small effect size (Akin-Little & Little, 2004). This study examined the overjustification effect and researchers selected students who already demonstrated high levels of rule following, which in turn resulted in a greater degree of overlap between the baseline and treatment phases. Excluding this one study, all the studies conducted in general education classrooms showed large effect sizes. For studies conducted in special education classrooms, one study (11%) showed a moderate effect size.

2.3 Discussion

This systematic review and meta-analysis included 24 token economy studies and reported the token economy components across general and special education classrooms ranging from kindergarten to 5th grade along with the effect sizes of each study. Results of the review confirm that different token economy components are used for different classroom types, and that token economy yields large effect sizes in classroom settings, regardless of type of classroom. Given that token economy interventions are highly effective in changing behavior, the differences in token economy components point to optimal arrangements for the respective classroom type and point to areas for future research.

2.3.1 Summary of Findings

General and special education classrooms used similar type of tokens and backup reinforcers. Points, small marks, or small items were most commonly used, and the tokens were small, portable, and easily accessible in a classroom setting possibly for broader use across different target behaviors and settings. Edibles and small tangibles were most used as backup
reinforcers. In a general education classroom, mystery rewards/motivators were second-most used while activities were more often used in a special education classroom. In a special education classroom, more concrete reinforcers such as edibles, tangibles, and activities were used, highlighting that students in special education classrooms may prefer more salient items or activities. The backup reinforcers used in general education classrooms were more communal while they were more individualized based on a student’s preference in special education classrooms.

Both classroom types mostly implemented a token economy system to address classroom behavior and academically related behaviors (e.g., following classroom rules, socially appropriate behavior, disruptive behavior, and on- and off-task behavior) rather than academic responding. The two studies that targeted academic skills such as reading (Alter, 2012; Shepley et al., 2016) were all conducted in special education classrooms and yielded large effect sizes (0.85 and 0.82, respectively).

Most token economies focused on positive reinforcement operations. Response cost (negative punishment operation) was more often used in a general education classroom compared to a special education classroom; however, across all classroom types, response cost was not often used as part of the independent variable. Interestingly, the two studies demonstrating moderate or small effect sizes (Akin-Little & Little, 2004; McDonald et al., 2014) did not use response cost, suggesting that token economies may be more effective when used in conjunction with response cost. The comparison would be stronger with a component analysis of token earn systems with and without response cost.

Regarding the token production rate, the studies in general education classrooms mostly measured instances of disruptive behavior while the studies in special education classrooms
mostly measured instances of correct target behavior. Further, the required time of engagement in appropriate behavior for each interval was longer in general education classrooms (range, 30 s to 50 min) than in special education classrooms (range, 20 s to 30 min). However, studies did not specify how the researchers determined the rate.

Across both classroom types, a fixed time (FT) exchange production rate was most commonly used, which was set at the end of the day or after a session. Special education classrooms, however, often used a fixed rate (FR) exchange production rate where a predetermined number of tokens was implemented. This may indicate that teachers in special education classrooms use a more concrete concept of number rather than the abstract concept of time to help students understand when they will be trading in. Regardless, in line with prior research, students in both classroom types preferred accumulated reinforcement (Ward-Horner, et al., 2017; Frank-Crawford et al., 2019), or larger reinforcers delivered at a later time. Overall, the time it took to obtain the backup reinforcer was greater in general education classrooms, which is not surprising if we assume greater self-control among general education populations. Future research should directly compare the effects FT and FR exchange production rate to determine which schedule is more effective in increasing or decreasing responses and yield greater maintenance of the behavior change.

Regarding token exchange rate, the number of tokens required was most commonly used across all settings. However, in general education classrooms, the second-most used token exchange rate incorporated positive punishment procedure. The system allowed the students to trade in if they earned a number of tokens below the criteria and the students were given a token when they displayed undesirable behavior. Notably, social competition was also often employed to change the students’ behavior as the students needed to win their peers or teachers to gain
access to reinforcers. The findings also showed that group contingencies were more often used in general education classrooms compared to special education classrooms. Only one study conducted in special education classrooms incorporated a group contingency (Groves & Austin, 2017). A recent meta-analysis showed that group contingencies were used in similar proportions across general and special education classrooms with large effect sizes (Little et al., 2015). The finding highlighted that group contingencies were in fact widely used across classroom types; however, group contingencies were less used in special education classrooms when used with token economies. There needs further research on the effects of a group contingency embedded in a token economy system in special education classrooms.

2.3.2 Implication for Future Practice

The findings suggested directions for future practice of token economies in each classroom type. Across both classroom types, many studies did not specify how they decided on a certain type of token or backup reinforcer. Formal preference assessments were not often reported and interventionists should strongly consider adding this element when implementing a token economy. Preference assessments allow better replicability and validity of the reinforcer selection (Piazza et al., 1996). Further, the findings showed that the majority of studies did not fade the time to access to tokens and/or backup reinforcers or response requirement over the course of the experiment. One study (Petursdottir & Ragnarsdottir, 2019) systematically faded the time it took to earn a token and number of tokens required to receive a backup reinforcer. The study showed large effect size, highlighting that fading would not negatively affect the effectiveness of a token economy. Thus, classrooms should develop token economy systems that systematically fade the time it takes to earn tokens and backup reinforcers to promote
generalization and self-control. Past studies showed that systematically increasing the duration to access reinforcers can also promote self-control (Dixon & Falcomata, 2004; Fisher et al., 2000).

In general education classrooms, token economies were narrowly applied to behavior management. However, the two studies that utilized token economies to teach academic skills in special education classrooms showed large effect sizes, and thus, educators of general education classrooms should be encouraged to implement token economies for academic responding. Also, studies that used group contingencies were highly effective in general education classrooms, highlighting that educators could explore different types or variations of group contingencies or peer-yoked contingencies to make the token economy system more effective. Similarly, studies that employed response cost had large effect sizes. This means that educators can incorporate response cost into token economy practices without limiting the effectiveness.

Mystery rewards/reinforcers were used less often in special education classrooms as compared to general education classrooms. However, the studies using mystery rewards/reinforcers in special education classrooms (Groves & Austin, 2017; Musser et al., 2001) had large effect sizes similar to the ones in general education classrooms (Mottram et al., 2002; Robichaux & Gresham, 2014). The finding indicated that special education classrooms may benefit from using mystery rewards/reinforcers and should consider employing mystery rewards/reinforcers more frequently. The finding also suggested that more naturalistic and social reinforcers should be employed in special education classrooms, assuming that those stimuli function to reinforce behavior. This is important because the behavior would come under the control of the natural contingencies and thus facilitate maintenance of the behavior. Similarly, studies using social competition in general education had large effect sizes, and thus educators in special education classrooms should explore the use of social competition in token economies.
Further, special education classrooms should be encouraged to incorporate group contingencies and response costs for token economies. The study with moderate effect size (McDonald et al., 2014) did not use response cost and the unit of contingency was independent. Hence, group contingencies and response cost may enhance the effectiveness of token economies in special education classrooms.

2.3.3 Implication for Future Research

The review suggests possible directions for future research. First, more studies should investigate the effects of token economies on academic instruction in general education classrooms. The two studies that utilized token economies to teach academic skills yielded large effect sizes, warranting the future use of token economies across various academic skills. Additional studies investigating the effects of token economy on academic responding would inform educators on how to use token economies when teaching academic skills and thus, broaden the scope of application for token economy systems in educational settings.

Second, there needs to be more research on how different types of behavior should produce tokens. This would require navigating the maximum response requirements students would be willing to engage without diminishing the perceived value of the reinforcer. Perhaps, making a decision tree to navigate the optimal token production rate depending on the response type or, in large, to arrange the token economy components should be encouraged. Additionally, as mentioned above, studies should develop token economy systems that systematically fade the time it takes to earn tokens and backup reinforcers to promote self-control. Further, a parametric analysis that compares different length of time it takes to obtain the backup reinforcer would provide more insight into an optimal exchange production rate.
Third, in extension to the point raised above, future research could develop a decision tree to determine optimal arrangements of token economy components based on settings and populations. A decision tree would function as a grid for educators and clinicians and allow a consistent method of token economy implementation, which would in turn encourage a consistent way of reporting token economies.

Fourth, researchers should study the effectiveness of token economies with mystery motivator/reinforcer and social competition in special education classrooms given that studies utilizing mystery motivator/reinforcer and social competition showed large effect sizes. Studies on mystery motivator/reinforcer and social competition in special education classrooms would help educators and clinicians effectively alter reinforcers and stimuli and discover novel ways to arrange token economy systems such as incorporating group and peer-yoked contingencies.

Fifth, a component analysis that compares token economies with and without response cost should be encouraged to further determine the effectiveness of response cost in token economies. To do this, researchers should vary the contingencies of the component analysis (e.g., reward and response cost) based on past research since studies showed that response cost affects different populations differently. For example, past studies showed that response cost improved math skills for students with Attention Deficit Hyperactivity Disorder (ADHD) compared to reward (Carlson et al., 2000).

Sixth, formal preference assessments to identify token types and backup reinforcers should be encouraged in token economy studies. This way, researchers can systematically replicate the procedures and increase validity in the use of the reinforcers.

Lastly, we encourage researchers to conduct more high-quality studies investigating the use of token economies in educational settings. Things to consider include but not limited to
having sufficient IOA for each condition, having more than three data points in each phase, and demonstrating an intervention effect across at least three points in time. A large number of studies excluded when the WWC standards were applied inform researchers that future studies should better adhere to the minimal standards to ensure high-quality single-case research in applied settings.

Despite strong empirical support for token economies, many of the studies described the token economy components vaguely described or did not specify them, which is consistent with the findings by Ivy et al. (2017). The problem arises in both research and practice. In research, the problem would be especially evident in systematic replications as pointed by Ivy et al. (2017). Systematic replications involve repeating the experiment with variations of the intervention variables (Sidman, 1960). To determine the effectiveness of individual intervention variables, systematic replications rely on sufficient descriptions of intervention components. Insufficient description of all components may make future systematic replications difficult. In practice, without a clear operational description of the essential components of token economies, the procedure may not be accurately implemented, affecting the effectiveness of the intervention. For best practice and replication of studies, universalizing the terms and how the procedure is written should be encouraged (Gutierrez et al., 2020).

2.3.4 Limitations

The review is not without limitations. First, identifying and categorizing the token economy components (i.e., token production rate, exchange production rate, and token exchange rate) were difficult due to the variability in how studies reported these components. This points to the importance of creating a general guideline on how token economy studies should report the procedure. Also, identifying components of a token economy system is a subjective process
and thus two reviewers may disagree whether a component was present or absent. To address this limitation, the researchers developed an operational definition for each component (Table 1).

Second, general education includes inclusive settings where students with special needs are part of the class with typically developing students (Szumski et al., 2017). While it continues to be an ambiguous term (Szumski et al., 2017), inclusive education has become more widely used and recommended since recent federal mandates holding the schools accountable for student achievement including students with disabilities (No Child Left Behind, 2002) and special education law requiring students with disabilities to have access to general education curriculum (Individuals with Disabilities Education Act, 2004). This means that the general education classroom studies identified in the current review may include inclusive settings, which could potentially affect the effectiveness of token economies. Future studies should determine whether token economies conducted in inclusive settings differ from noninclusive general education classrooms to better inform educators on token economy use in these types of classrooms.

2.3.5 Conclusion

The purpose of this meta-analysis was to determine the effectiveness of token economy systems and which token economy component variations make the system most effective across general and special education classrooms across grade-levels from kindergarten to 5th grade. The results showed that token economies yield large effect sizes in general and special education classroom types with slightly larger average effect size in special education classrooms compared to general education classrooms. The results also demonstrated that there are variations in the token economy components depending on the classroom type, which shed light on the optimal use of token economies in each respective classroom. The findings suggest
directions for future directions such as studies that directly compare component variations and consistent way of reporting the critical components for future replication and application.

Identification of key components of token economy systems in different classroom settings may serve as the first step in developing an optimal token economy system across many settings and populations.
References

An * (asterisks) indicates articles included in the systematic review and meta-analysis


https://doi.org/10.1023/B:JOBE.0000037628.81867.69


https://doi.org/10.1080/1045988X.2011.565283


https://doi.org/10.1901/jaba.1968.1-13


https://doi.org/10.1002/pits.21858


https://doi.org/10.1207/S15328023TOP3003_05


Corrective and Social Psychiatry and Journal of Behavior Technology Methods and Therapy, 27(3), 140-145.


Table 1

**Definitions of article analysis components**

<table>
<thead>
<tr>
<th>Token economy component</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of tokens</td>
<td>The description of the type of tokens used to reinforce the target behavior.</td>
</tr>
<tr>
<td>Types of backup reinforcers</td>
<td>The description of the type of backup reinforcers used to exchange with the tokens.</td>
</tr>
<tr>
<td>Target behavior</td>
<td>The description of the response class producing a token.</td>
</tr>
<tr>
<td>Response cost</td>
<td>The response-contingent loss of a specific number of positive reinforcer; a fine</td>
</tr>
<tr>
<td>Token production rate</td>
<td>The rate at which delivery of tokens are made respective to the number of target behaviors emitted; response requirements.</td>
</tr>
<tr>
<td>Exchange-production rate</td>
<td>The rate at which tokens can be exchanged for backup reinforcers.</td>
</tr>
<tr>
<td>Token-exchange rate</td>
<td>The cost of back up reinforcers in terms of tokens.</td>
</tr>
<tr>
<td>Unit of contingency</td>
<td>Whether access to the backup reinforcer depends on the performance of a group as a whole or an individual. The units are called group or independent, respectively.</td>
</tr>
</tbody>
</table>

*Note.* The definitions were adapted from Hackenberg (2009), Ivy et al. (2017), Kazdin and Bootzin (1972), and Reed et al. (2013).
<table>
<thead>
<tr>
<th>Journal</th>
<th>Number of Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral Development Bulletin</td>
<td>1</td>
</tr>
<tr>
<td>Behavioral Interventions</td>
<td>1</td>
</tr>
<tr>
<td>Education and Training in Autism and Developmental Disabilities</td>
<td>1</td>
</tr>
<tr>
<td>Education and Treatment of Children</td>
<td>2</td>
</tr>
<tr>
<td>Electronic Journal of Research in Educational Psychology</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Applied Behavior Analysis</td>
<td>3</td>
</tr>
<tr>
<td>Journal of Applied School Psychology</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Behavioral Education</td>
<td>4</td>
</tr>
<tr>
<td>Journal of Deaf Studies and Deaf Education</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Positive Behavior Interventions</td>
<td>1</td>
</tr>
<tr>
<td>Journal of School Psychology</td>
<td>1</td>
</tr>
<tr>
<td>School Psychology Review</td>
<td>2</td>
</tr>
<tr>
<td>Preventing School Failure</td>
<td>1</td>
</tr>
<tr>
<td>Psychology in the Schools</td>
<td>3</td>
</tr>
<tr>
<td>Remedial and Special Education</td>
<td>1</td>
</tr>
<tr>
<td>Reference</td>
<td>Tokens</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Akin-Little &amp; Little (2004)</td>
<td>Points</td>
</tr>
<tr>
<td>Battaglia et al. (2015)</td>
<td>&quot;X&quot; marks</td>
</tr>
<tr>
<td>Christensen et al. (2004)</td>
<td>Plastic chips</td>
</tr>
<tr>
<td>Christensen et al. (2007)</td>
<td>Points and plastic coins</td>
</tr>
<tr>
<td>De Martini-Scully et al. (2000)</td>
<td>Not specified</td>
</tr>
<tr>
<td>Reference</td>
<td>Tokens</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Donaldson et al. (2014)</td>
<td>Check marks</td>
</tr>
<tr>
<td>Lastrapes et al. (2018)</td>
<td>Points</td>
</tr>
<tr>
<td>Ling et al. (2011)</td>
<td>Smiley face</td>
</tr>
<tr>
<td>Lynne et al. (2017)</td>
<td>ClassDojo points</td>
</tr>
<tr>
<td>Mottram et al. (2002)</td>
<td>Not specified</td>
</tr>
<tr>
<td>Pennington &amp; McComas (2017)</td>
<td>Points</td>
</tr>
<tr>
<td>Reference</td>
<td>Tokens</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Petursdottir &amp; Ragnarsdottir (2019)</td>
<td>Star stamps</td>
</tr>
<tr>
<td>Robichaux &amp; Gresham (2014)</td>
<td>Tally marks</td>
</tr>
<tr>
<td>Tanol et al. (2010)</td>
<td>Stars</td>
</tr>
<tr>
<td>Wright &amp; McCurdy (2011)</td>
<td>Points</td>
</tr>
</tbody>
</table>

*Note. FI = Fixed Interval schedule of reinforcement; FR = Fixed Ratio schedule of reinforcement; VI = Variable Interval schedule of Reinforcement; VR = Variable Ratio schedule of reinforcement.*
<table>
<thead>
<tr>
<th>Reference</th>
<th>Tokens</th>
<th>Backup reinforcers</th>
<th>Target behavior</th>
<th>Response cost</th>
<th>Token production</th>
<th>Exchange production</th>
<th>Token exchange</th>
<th>Unit of contingency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alter (2012)</td>
<td>Points</td>
<td>Activities, tangibles, and edibles</td>
<td>Math word problem accuracy</td>
<td>N</td>
<td>Completion of each step</td>
<td>End of 15-min session</td>
<td>Not specified</td>
<td>Independent</td>
</tr>
<tr>
<td>Carnett et al. (2014)</td>
<td>Pennies, small foam puzzle pieces</td>
<td>Edibles</td>
<td>Disruptive and on-task behavior</td>
<td>N</td>
<td>FI 20 s of on-task behavior</td>
<td>FR 10</td>
<td>FR 10</td>
<td>Independent</td>
</tr>
<tr>
<td>Gann et al. (2015)</td>
<td>Penny and not specified for two participants</td>
<td>Not specified</td>
<td>Task engagement</td>
<td>N</td>
<td>FR 1 of on-task behavior</td>
<td>Not specified</td>
<td>Not specified</td>
<td>Independent</td>
</tr>
<tr>
<td>Groves &amp; Austin (2017)</td>
<td>Points</td>
<td>Mystery prize (activities and edibles)</td>
<td>Disruptive and off-task behavior</td>
<td>N</td>
<td>FI 2 min of on-task behavior and following rules</td>
<td>Immediately after game</td>
<td>Mystery number from 1 -25</td>
<td>Group in one condition</td>
</tr>
<tr>
<td>McDonald, Reeve, &amp; Sparacio (2014)</td>
<td>Pennies</td>
<td>Edibles, tangibles, and activities</td>
<td>Stereotypic behavior</td>
<td>N</td>
<td>FI 10 min of incompatible behavior of stereotypy</td>
<td>Not specified</td>
<td>FR 10</td>
<td>Independent</td>
</tr>
<tr>
<td>Reference</td>
<td>Tokens</td>
<td>Backup reinforcers</td>
<td>Target behavior</td>
<td>Response cost</td>
<td>Token production</td>
<td>Exchange production</td>
<td>Token exchange</td>
<td>Unit of contingency</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------</td>
<td>---------------------------</td>
<td>-----------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>----------------------</td>
<td>-----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Musser et al. (2001)</td>
<td>Stickers</td>
<td>Mystery motivator</td>
<td>Disruptive behavior</td>
<td>Y</td>
<td>FI 30 min of compliance</td>
<td>FI 8 thinned to 10 and 12</td>
<td>FI 8 thinned to 10 and 12</td>
<td>Independent</td>
</tr>
<tr>
<td>Shepley et al. (2016)</td>
<td>Yellow star</td>
<td>Not specified (student records and teacher suggestion)</td>
<td>Reading accuracy</td>
<td>N</td>
<td>FR 1 of prompted or unprompted correct responding</td>
<td>FR 10</td>
<td>FR 10</td>
<td>Independent</td>
</tr>
<tr>
<td>Thompson et al. (2011)</td>
<td>Tickets</td>
<td>Tangibles</td>
<td>Disruptive behavior</td>
<td>N</td>
<td>Improved by at least one talk-out than previous session and/or one or no talk-outs</td>
<td>FR 5</td>
<td>FR 5</td>
<td>Independent</td>
</tr>
<tr>
<td>Wadsworth et al. (2015)</td>
<td>Smiley face stickers</td>
<td>Activities, tangibles, and edibles</td>
<td>Compliance</td>
<td>N</td>
<td>FR 1 of compliance</td>
<td>End of session</td>
<td>FR 4</td>
<td>Independent</td>
</tr>
</tbody>
</table>

*Note.* FI = Fixed Interval schedule of reinforcement; FR = Fixed Ratio schedule of reinforcement.
Figure 1

Schematic overview of the data search and data analysis process

Stage 1: Keyword Search

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Number of Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Token Economy”</td>
<td>563</td>
</tr>
<tr>
<td>“Token Reinforcement”</td>
<td>256</td>
</tr>
<tr>
<td>Total: 819</td>
<td></td>
</tr>
</tbody>
</table>

Stage 2: First Level of Analysis
176 articles

Stage 2: Second Level of Analysis
20 included

Stage 3: Citation Search
1,064 hits

Stage 3: Citation Search
15 included (novel)

Stage 3: Reference Search
571 hits

Stage 3: Reference Search
4 included (novel)

39 articles identified

WWC Standards Applied

Total
24 included
Figure 2

Effect Size Using Improvement Rate Difference

<table>
<thead>
<tr>
<th>General Education</th>
<th>Small Effect</th>
<th>Moderate Effect</th>
<th>Large Effect</th>
<th>IRD [CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akin-Little &amp; Little (2004)</td>
<td>0.05 [0.04, 0.15]</td>
<td>0.93 [0.65, 0.96]</td>
<td>0.91 [0.65, 0.97]</td>
<td></td>
</tr>
<tr>
<td>Battaglia et al. (2015)</td>
<td>1 [0.7, 1]</td>
<td>0.76 [0.6, 0.84]</td>
<td>0.83 [0.74, 0.9]</td>
<td></td>
</tr>
<tr>
<td>Christensen et al. (2004)</td>
<td>0.72 [0.64, 0.79]</td>
<td>1 [0.72, 1]</td>
<td>0.8 [0.66, 0.87]</td>
<td></td>
</tr>
<tr>
<td>Christensen et al. (2007)</td>
<td>0.95 [0.81, 0.97]</td>
<td>0.84 [0.69, 0.91]</td>
<td>0.83 [0.49, 0.95]</td>
<td></td>
</tr>
<tr>
<td>De Martinis-Scully et al. (2000)</td>
<td>0.89 [0.69, 0.96]</td>
<td>0.97 [0.85, 0.98]</td>
<td>0.93 [0.78, 0.96]</td>
<td></td>
</tr>
<tr>
<td>Donladson et al. (2014)</td>
<td>0.83 [0.49, 0.95]</td>
<td>0.89 [0.69, 0.96]</td>
<td>0.97 [0.85, 0.98]</td>
<td></td>
</tr>
<tr>
<td>Larriaga et al. (2018)</td>
<td>0.93 [0.78, 0.96]</td>
<td>0.83 [0.49, 0.95]</td>
<td>0.89 [0.69, 0.96]</td>
<td></td>
</tr>
<tr>
<td>Ling et al. (2011)</td>
<td>1 [0.85, 1]</td>
<td>0.87</td>
<td>0.85 [0.65, 0.94]</td>
<td></td>
</tr>
<tr>
<td>Lynne et al. (2017)</td>
<td>0.83 [0.49, 0.95]</td>
<td>0.89 [0.69, 0.96]</td>
<td>0.97 [0.85, 0.98]</td>
<td></td>
</tr>
<tr>
<td>Mottram et al. (2002)</td>
<td>0.83 [0.49, 0.95]</td>
<td>0.89 [0.69, 0.96]</td>
<td>0.97 [0.85, 0.98]</td>
<td></td>
</tr>
<tr>
<td>Pennington &amp; McComas (2017)</td>
<td>1 [0.83, 1]</td>
<td>0.91 [0.74, 0.95]</td>
<td>0.91 [0.74, 0.95]</td>
<td></td>
</tr>
<tr>
<td>Petursdottir &amp; Ragnarsdottir (2019)</td>
<td>0.81 [0.52, 0.81]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.78, 0.92]</td>
<td></td>
</tr>
<tr>
<td>Robichaud &amp; Grebusha (2014)</td>
<td>0.82 [0.78, 0.92]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.56, 0.9]</td>
<td></td>
</tr>
<tr>
<td>Tatom et al. (2010)</td>
<td>0.87</td>
<td>0.82 [0.56, 0.9]</td>
<td>1 [0.85, 1]</td>
<td></td>
</tr>
<tr>
<td>Wright &amp; McCurdy (2011)</td>
<td>0.82 [0.56, 0.9]</td>
<td>1 [0.85, 1]</td>
<td>0.87</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Special Education</th>
<th>Small Effect</th>
<th>Moderate Effect</th>
<th>Large Effect</th>
<th>IRD [CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alter (2012)</td>
<td>0.83</td>
<td>0.85 [0.65, 0.94]</td>
<td>0.93 [0.78, 0.96]</td>
<td></td>
</tr>
<tr>
<td>Carnett et al. (2014)</td>
<td>0.81 [0.52, 0.81]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.78, 0.92]</td>
<td></td>
</tr>
<tr>
<td>Gann et al. (2015)</td>
<td>0.91 [0.74, 0.95]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.56, 0.9]</td>
<td></td>
</tr>
<tr>
<td>Groves &amp; Austin (2017)</td>
<td>0.81 [0.52, 0.81]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.56, 0.9]</td>
<td></td>
</tr>
<tr>
<td>McDonald et al. (2014)</td>
<td>0.82 [0.78, 0.92]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.56, 0.9]</td>
<td></td>
</tr>
<tr>
<td>Musser et al. (2001)</td>
<td>0.82 [0.78, 0.92]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.56, 0.9]</td>
<td></td>
</tr>
<tr>
<td>Shipley et al. (2016)</td>
<td>0.87</td>
<td>0.82 [0.56, 0.9]</td>
<td>1 [0.85, 1]</td>
<td></td>
</tr>
<tr>
<td>Thompson et al. (2011)</td>
<td>0.81 [0.52, 0.81]</td>
<td>0.94 [0.72, 0.99]</td>
<td>0.82 [0.56, 0.9]</td>
<td></td>
</tr>
<tr>
<td>Wadsworth et al. (2015)</td>
<td>0.87</td>
<td>0.82 [0.56, 0.9]</td>
<td>1 [0.85, 1]</td>
<td></td>
</tr>
<tr>
<td>Sp Ed Average</td>
<td>0.87</td>
<td>0.82 [0.56, 0.9]</td>
<td>1 [0.85, 1]</td>
<td></td>
</tr>
</tbody>
</table>

Note. IRD = improvement rate difference; CI = confidence interval.
## Appendix

### 2009 PRISMA Checklist

*The page numbers correspond to the numbers on the published manuscript.

<table>
<thead>
<tr>
<th>Section/topic</th>
<th>Checklist item</th>
<th>Reported on page #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TITLE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Identify the report as a systematic review, meta-analysis, or both.</td>
<td>1</td>
</tr>
<tr>
<td><strong>ABSTRACT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structured summary</td>
<td>Provide a structured summary including, as applicable: background, objectives, data sources, study eligibility criteria, participants, and interventions, study appraisal and synthesis methods, results, limitations, conclusions and implications of key findings, systematic review registration number.</td>
<td>2</td>
</tr>
<tr>
<td><strong>INTRODUCTION</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale</td>
<td>Describe the rationale for the review in the context of what is already known.</td>
<td>4-6</td>
</tr>
<tr>
<td>Objectives</td>
<td>Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).</td>
<td>6</td>
</tr>
<tr>
<td><strong>METHODS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocol and registration</td>
<td>Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.</td>
<td>NA</td>
</tr>
<tr>
<td>Eligibility criteria</td>
<td>Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.</td>
<td>7-8</td>
</tr>
<tr>
<td>Information sources</td>
<td>Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.</td>
<td>7-8</td>
</tr>
<tr>
<td>Search</td>
<td>Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.</td>
<td>7-8</td>
</tr>
<tr>
<td>Study selection</td>
<td>State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).</td>
<td>7-8</td>
</tr>
<tr>
<td>Data collection process</td>
<td>Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.</td>
<td>7-9</td>
</tr>
<tr>
<td>Data items</td>
<td>List and define all variables for which data were sought (e.g., PICOS, funding sources) and any processes for obtaining and confirming data from investigators.</td>
<td>9-10</td>
</tr>
<tr>
<td>Risk of bias in individual studies</td>
<td>Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.</td>
<td>9</td>
</tr>
<tr>
<td>Summary measures</td>
<td>State the principal summary measures (e.g., risk ratio, difference in means).</td>
<td>10-11</td>
</tr>
<tr>
<td>Synthesis of results</td>
<td>Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I²) for each meta-analysis.</td>
<td>10-11</td>
</tr>
<tr>
<td>Section/topic</td>
<td>#</td>
<td>Checklist item</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Risk of bias across studies</td>
<td>15</td>
<td>Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).</td>
</tr>
<tr>
<td>Additional analyses</td>
<td>16</td>
<td>Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.</td>
</tr>
<tr>
<td><strong>RESULTS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study selection</td>
<td>17</td>
<td>Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.</td>
</tr>
<tr>
<td>Study characteristics</td>
<td>18</td>
<td>For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.</td>
</tr>
<tr>
<td>Risk of bias within studies</td>
<td>19</td>
<td>Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).</td>
</tr>
<tr>
<td>Results of individual studies</td>
<td>20</td>
<td>For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.</td>
</tr>
<tr>
<td>Synthesis of results</td>
<td>21</td>
<td>Present results of each meta-analysis done, including confidence intervals and measures of consistency.</td>
</tr>
<tr>
<td>Risk of bias across studies</td>
<td>22</td>
<td>Present results of any assessment of risk of bias across studies (see item 15).</td>
</tr>
<tr>
<td>Additional analysis</td>
<td>23</td>
<td>Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see item 16]).</td>
</tr>
<tr>
<td><strong>DISCUSSION</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary of evidence</td>
<td>24</td>
<td>Summarize the main findings including the strength of evidence for each main outcome, consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).</td>
</tr>
<tr>
<td>Limitations</td>
<td>25</td>
<td>Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).</td>
</tr>
<tr>
<td>Conclusions</td>
<td>26</td>
<td>Provide a general interpretation of the results in the context of other evidence, and implications for future research.</td>
</tr>
<tr>
<td><strong>FUNDING</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Funding</td>
<td>27</td>
<td>Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.</td>
</tr>
</tbody>
</table>
Chapter 3: A Rapid Assessment of Sensitivity to Reward Delays and Classwide Token Economy Savings for School-Aged Children

Outline

Delay discounting tasks measure the relation between reinforcer delay and efficacy. The present study introduced a brief, user-friendly measure for educators interested in quantifying sensitivity to reward delays for school-aged children. Study 1 reanalyzed data collected by Reed and Martens (2011) and found that 1-month delay choices predicted student classroom behavior. Study 2 examined the utility of the 1-month delay indifference point in predicting saving and spending behavior of second-grade students using token economies with two different token production schedules. Collectively, results showed (a) the 1-month delay indifference point predicted classroom behavior, (b) children who discounted less and had greater self-regulation, accrued and saved more tokens, (c) a variable token production schedule better correlated with discounting than a fixed schedule, and (d) there were group differences based on child characteristics regarding sensitivity to token production schedule change. Implications are discussed regarding utility of a rapid discounting assessment for educational settings.

Keywords: delay discounting, classroom, sensitivity to reward delays, token economy.
A Rapid Assessment of Sensitivity to Reward Delays and Classwide Token Economy Savings for School-Aged Children

Delay discounting is a concept within behavioral economics that examines preferences between smaller/sooner and larger/later reinforcers (abbreviated as SSR and LLR, respectively). Accordingly, this concept presents applied implications for constructs such as “delay of gratification” (Mischel, 1983) or tolerance for delayed reinforcers. While behavioral economics researchers study the effects of reinforcer constraints on operant responses (Hursh & Roma, 2016)—typically measured in operant demand—behavioral economists have translated microeconomic concepts to study delay discounting. Delay discounting, defined as the systematic devaluation of reinforcement as a function of increasing temporal delays, quantifies an organism’s operant sensitivity to delayed reinforcers. The delay discounting process has been studied extensively in both human and nonhuman studies (e.g., Vanderveldt et al., 2016), using a variety of procedures, ranging from hypothetical choice arrangements to experiential tasks (see review in Reed et al., 2020); central to all discounting tasks is the measurement of organisms’ responding towards either SSRs or LLRs.

Despite the apparent translational utility of delay discounting in bridging applied behavior analytic concepts and broader constructs outside behaviorism, this topic remains relatively under-researched within applied behavior analysis. For example, of the 1,200 articles indexed through the year 2019 in Web of Science using the author keyword “delay discounting,” only 9 appear in the pages of Journal of Applied Behavior Analysis, despite delay discounting emanating from behavior analytic research (e.g., Rachlin et al., 1991). Quantifying sensitivity to delayed reinforcement renders it especially conducive to understanding everyday behavior related to constructs such as impulsivity (for a critique of this term, though, see Strickland &
Johnson, 2020), which are thought to be associated with myopic reinforcer preferences. Indeed, delay discounting is associated with many maladaptive outcomes of social importance (e.g., myriad health problems [Daugherty & Brase, 2010], psychiatric conditions [Amlung et al., 2019] cigarette smoking [Mitchell, 1999], drug and alcohol abuse [Richards et al., 1999], overeating [Weller et al., 2008], and gambling [Reynolds, 2006]). Shallow discounting is also associated with “delay of gratification,” or what some behavior scientists call “self-control,” which is the behavior of selecting the more advantageous LLR over a less advantageous SSR (Dixon & Tibbetts, 2009). Behavior analysts have thereby proposed that delay discounting may be a promising approach to understanding socially important behavior from an operant perspective, while using concepts and procedures more readily accepted and understood by mainstream psychology (see Critchfield & Kollins, 2001).

Outside of applied behavior analysis, the majority of research on delay discounting has been conducted in laboratory settings with adult populations completing hypothetical delay discounting tasks. Nevertheless, studies have shown that younger populations discount delayed rewards (Reed & Martens, 2011) and that discounting relates to real-world outcomes that are of interests to educators and clinicians. Researchers have shown that adolescents discount more compared to adults (Steinberg et al., 2009) and that greater delay discounting correlated with lower IQ (Dougherty et al., 2014) and lower educational attainment (Jaroni et al., 2004). A review of delay discounting studies with participants with a mean age of 12.99 years old or younger showed: (a) the majority of studies used hypothetical choices about money, (b) less than half of the assessments included visual aids for comprehension of choice options, and (c) only one study was conducted in a classroom setting (Staubitz et al., 2018).
In an applied behavior analytic example, Reed and Martens (2011) investigated delay discounting in a classroom setting and compared reinforcement delivered immediately after class or in 24 hours when the students could exchange for backup reinforcers. The researchers found that for students with higher discounting rates, the delayed rewards were less effective in improving on-task behavior compared to immediate rewards. Given that younger populations discount delayed rewards, additional studies on delay discounting in an applied setting with younger populations are needed to translate the laboratory measures of delay discounting to a more naturalistic setting. The findings can help educators find the optimal use of rewards and schedules of reinforcement to promote delay of gratification in younger populations especially those at higher risk for poor long-term outcomes.

Further, delay discounting has clear implications for saving and spending behavior. People set a financial goal but often divert from their initial plan and spend money on a smaller, immediately available goods or services at the expense of saving money for a more advantageous, delayed reward. One way to measure children’s saving and spending behavior is by using a classwide token economy system which mimics a real-world economy where consumers (i.e., students) and producers (i.e., teachers) interact to allocate limited resources (i.e., backup reinforcers). Students’ saving and spending behavior can be captured in this process, and researchers can thus use the data to relate it to delay discounting measure. A classwide token economy captures younger students’ spending and saving behavior because these behaviors start to evolve at a young age as children develop the ability to grasp the concept of “temptation” and the consequences of spending sooner on future opportunities (Sonuga-Barke & Webley, 1993). Children between ages of 6 and 12 years start to understand the value of saving and the association between saving and better future opportunity (Te’eni-Harari, 2016). Around 12 years
old, young individuals formulate and engage in more complex saving and spending strategies (Sherraden, 2011). Studies also found that saving behavior in adolescence was associated with adult saving (Ashby et al., 2011). In general, older children save more money compared to younger children (Mischel & Mischel, 1983). A recent study showed that delay discounting is related to risky financial and household savings behavior for smokers (Snider et al., 2019); however, more studies are needed to understand the relation between delay discounting and saving and spending behavior of younger populations. Understanding children’s saving and spending behavior and relating it to delay discounting can broaden the scope of delay discounting and provide insight into its utility with children, which would inform researchers on the development of interventions aimed to improve long-term delay of gratification and financial decision-making.

Despite the applicability of delay discounting measures across populations and behavioral problems and implications for saving and spending behavior, the complex nature of deriving meaningful numbers impedes its widespread use among educators and clinicians (Critchfield & Reed, 2009). Typically, researchers would find several indifference points, which are points at which the value switches from the LLR to the SSR (i.e., the points at which the subjective values of LLR and SSR are equal; Reed et al., 2013), through delay discounting tasks with binary intertemporal choices and use them to find discounting parameters (e.g., $k$ values; see Odum, 2011) through nonlinear modeling (see Gilroy et al., 2017) and area under the discounting curve (AUC; e.g., Myerson et al., 2001) to quantify sensitivity to reward delays. Finding the parameters and AUC involves mathematical equations and a certain level of knowledge in statistics and software (Motulsky & Christopoulos, 2006). A development of a more accessible,
user-friendly method would lower the barrier for educators to evaluate the degree of delayed reward sensitivity of younger populations.

In response, the current study investigated a brief intertemporal choice task for applied settings that can be used to assess the degree of delayed reward sensitivity of younger populations. The original study plan included investigating the relation between student saving and spending behavior using a classwide token economy system and delay discounting measures with several indifference points. However, when schools closed as a response to the COVID-19 pandemic during the course of the study, the researchers were only able to complete the delay discounting task at 1-month delay and serendipitously found that the indifference point at 1-month delay may be enough to predict token economy saving and spending behavior for a younger population. Unexpected lines of research and findings like such refer to “discovery research” (Roane et al., 2003) and the serendipitous findings can be meaningful and noteworthy for the scientific community. Therefore, the primary goal of this study was to bridge the gap between basic and applied research on delay discounting by (a) reanalyzing the discounting task data from Reed and Martens (2011; Study 1 of this report), and (b) introducing a rapid assessment called the Brief Intertemporal Choice Task (BRIC task) to assess the degree of sensitivity to delayed rewards in a younger population and measure how performance on this task predicts student saving and spending behavior (Study 2 of this report). A secondary goal was to determine group differences based on child characteristics in saving and spending patterns across the different token production schedules for school-aged children and to examine whether individual differences in children’s delay discounting were related to another measure of self-regulation behavior.

3.1 STUDY 1
The purpose of Study 1 was to reanalyze discounting task data from Reed and Martens (2011) to evaluate (a) test-retest reliability of indifference points associated with the 1 month delay value used in their assessment for all 46 participants in the classwide intervention (the published study only included discounting data for the 26 students whose data met inclusionary criteria for systematic discounting and were analyzed with nonlinear curve-fitting), (b) correlations between the 1 month delay indifference point and AUC of the overall discounting function, (c) analysis of ED50 of the $k$ values from the 26 participants in the original study, and (d) correlations (for all 46 participants) between the 1 month delay indifference point and their primary dependent variable from the classwide token system: the difference in mean on-task behavior between immediate and delayed token exchanges.

3.1.1 Method

3.1.1.1 Discounting Data

As described in Reed and Martens (2011), the original study assessed delay discounting and on-task behavior for 46 participants who averaged 12.10 ($SD = .30$) years of age. The discounting procedure consisted of a rapid titration of binary choices between SSRs and LLRs (monetary; the magnitude of the undiscounted LLR was $100$) across delays ranging from 1 day to 4 years. The researchers administered the discounting procedure to each participant twice, with each procedure separated by 1 week.

Delay discounting analyses in the original study were limited to the 26 participants meeting the following inclusionary criteria: “if the mean of the indifference points from the three shortest delay conditions exceeded the mean of the indifference points from the three longest delay conditions, with no more than … two instances of increasing indifference points across successive delays” (p. 8). The researchers fitted both the hyperbolic (Mazur, 1987) and
hyperboloid discounting (Myerson & Green, 1995) models to render rates of discounting (i.e., \( k \) values) to these 26 participants’ discounting. However, all 46 participants returned indifference points for all delays; thus, AUC (see original study) was available for all 46 participants, as well.

### 3.1.1.2 On-Task Behavior Direct Observation Data

The Reed and Martens study (2011) included direct observation of on-task behavior for all 46 participants under both immediate (following the 20-min observation) and delayed (24-hrs after observation) token exchange delays during a classwide intervention (see original study for specific details). The authors calculated the difference in off-task behavior between the two exchange delay phases of the intervention as a primary dependent variable.

### 3.1.2 Results and Discussion

Table 1 displays a summary of the findings of the reanalysis. Reanalysis of the 46 participants’ discounting data from the original Reed and Martens study (2011; i.e., data not reported in their original study) suggested the indifference points and AUC values did not pass the D’Agostino and Pearson normality test, indicating the need for nonparametric statistics. Further reanalysis yielded a median 1-month indifference point of $30.00 (IQR = 7.00 to 70.00) at Time 1, and a median of $50.00 (IQR = 6.25 to 80.00) at Time 2. The median AUC was 0.13 (IQR = 0.04 to 0.24) at Time 1, and a median of 0.12 (IQR = 0.02 to 0.29) at Time 2. The test-retest reliability of the 1-month indifference point was significantly strong, \( r_s = 0.80, p < .001, N = 46 \). Note that the reported \( r_s \) for the 26 participants meeting criteria for systematic discounting in the original study was 0.81; thus, the data from all 46 participants is as reliable as those for the 26 participants subsample in the original study. Similarly, the test-retest reliability of AUC was significant and strong, \( r_s = .83, p < .001, N = 46 \). Finally, correlations between the 1-month indifference point and AUC at was significant and relatively strong at both Time 1 (\( r_s = .45, p \)
< .001, N = 46) and Time 2 (rs = .58, p < .001, N = 46). In sum, it appears the 1-month indifference point is reliable across 1 week, and is a sufficient proxy to an overall AUC.

Reed and Martens (2011) provided k values for the 26 participants meeting inclusion criteria for fitting discounting curves. At the time of their publication, they failed to explore a secondary analysis of discounting rate: ED50. Yoon and Higgins (2008) proposed the inverse of k renders the “effective delay” (ED) associated with 50% discounting—this ED50 variable is convenient as it provides a simple delay-scaled metric associated with discounting. That is, ED50 permits a precise quantification of the delay (in time-based units like days/weeks) associated with a 50% reduction in the subjective value of delayed reinforcer, which provides an easily communicable metric of delayed reward sensitivity to non-behaviorists. Across the 26 participants’ data from Table 1 of Reed and Martens, ED50 was calculated for the 20 participants with non-zero k values; ED50 was normally distributed (according to the D’Agostino & Pearson test) and ranged from 0.54 days to 100 days, with the mean ED50 equal to 34.27 days (SD = 40.14)—approximately equal to 1 month. That ED50 approximated 1 month lends further evidence that the 1-month delay is particularly useful in understanding school-aged students’ discounting and sensitivity to reward delays.

To evaluate the potential of a one-time indifference point at the 1-month delay serving as a proxy to sensitivity to reward delay in practice, researchers examined the correlation between the Time 1 indifference points and the difference scores between immediate and delayed classroom contingencies. The correlation was significant and moderate, rs = -0.33, p = .026, N = 46, suggesting that higher indifference points (i.e., higher valuation of a reward at 1-month delay) were associated with smaller sensitivity to classroom token delay contingencies. Simply put, participants who featured less differences in on-task behavior between immediate and
delayed token exchange schedules also featured higher indifference points. Collectively, these data suggest that sensitivity to hypothetical 1-month monetary reward delays are significantly associated with differential sensitivity in actual operant contingencies in the classroom.

In sum, the findings of this reanalysis suggest the 1-month delay from Reed and Martens’ discounting task (2011) appear to be reliable and predictive of classroom behavior. These findings lend promise to the potential use of a single delay discounting task as a rapid assessment of reinforcement delay sensitivity in school-aged children.

3.2 STUDY 2

The reanalysis of Reed and Martens (2011) showed that the indifference point at 1-month period significantly related to actual classroom behavior and adequately proxied participants’ sensitivity to reward delay. The primary purpose of Study 2 was to extend the findings of Study 1 by (a) examining how the 1-month delay from the discounting task predicts school-aged children’s spending and saving behavior in a classwide token economy, and (b) establishing a rapid assessment of reinforcement delay sensitivity in school-aged children, also called the Brief Intertemporal Choice Task (BRIC Task). The secondary purpose was to study group differences for responding to two different token production schedules. The researchers studied the relation between the indifference points obtained by the BRIC Task and two different token production schedules. Finally, because laboratory measures of delay of gratification have been linked to children’s self-regulation, or the capacity to resist a prepotent response (e.g., Jahromi et al., 2019), the researchers also investigated whether delayed discounting was related to individual differences in teachers’ ratings of children’s self-regulation on a measure often used in mainstream psychology, the Behavior Rating Inventory of Executive Functioning 2 (BRIEF2; Gioia et al., 2015).
3.2.1 Method

3.2.1.1 Participants

Seventeen second graders (7 female and 10 male participants) participated. The participants’ ages ranged from 7.08 to 7.92 (M = 7.54) years old. Four participants had Individualized Education Programs (IEP) with classifications of specific learning disability, autism spectrum disorder, language impairment, and other health impairment with the diagnosis of Attention-Deficit/Hyperactivity Disorder (ADHD). One participant qualified for a 504 plan. Of the 17 participants, seven participants received free/reduced lunch (i.e., an index of a participant’s family’s socio-economic status), and six participants qualified as English Language Learners (ELL). Fifteen participants had experience with a classwide token economy system the prior school year. Two participants were new to the token economy system and the researcher conditioned tokens in the beginning of the year by pairing tokens with verbal praise and preferred activities and objects. One female participant who was both an ELL and received free/reduced lunch moved schools in the middle of the year. For this participant, we collected data for the first half of the study (i.e., fixed token production schedule and Behavior Rating Inventory of Executive Functioning 2 (BRIEF2) assessment) but did not for the second half of the study (i.e., the variable token production schedule and hypothetical delay discounting assessment). Put simply, we included 17 participants for the fixed token production schedule condition and BRIEF2 (i.e., secondary analysis) but 16 participants for the variable token production schedule condition and the BRIC Task (i.e., primary analysis). Prior to the study, all participants mastered learning money concepts involving bills and coins during math per the second-grade common core standard (2.MD.C.8) and mastered counting to 120 and comparing
two two-digit numbers per the first-grade common core standard (1.NBT.A.1 and 1.NBT.B.3, respectively).

### 3.2.1.2 Setting and Materials

The study took place in a public elementary school located in a suburb outside of a large metropolitan area with grades from pre-school to second. All participants were in a second-grade classroom that used a Comprehensive Application for Behavior Analysis to Schooling (CABAS®, Accelerated Independent Learner program; Greer, 1994) model with 18 students, one head teacher, and two teaching assistants. The model incorporates a scientific approach to pedagogy, learning, curriculum, and classroom management, and all instruction is individualized to each student’s repertoire. All token economy sessions took place in the participants’ classroom throughout the day and trade-in took place at the end of each school day. The researchers conducted the BRIC Task at individual tables in the classroom during the school day and teachers completed the BRIEF2 during the same month other measures were collected.

Materials for the fixed and variable token production schedule conditions included various tangible prizes for the prize store (i.e., 16 different types of small toys and school supplies), seven plastic prize boxes with numbers 1, 3, 5, 10, 20, 50, and 100 attached, paper money called “criterion cash” with a picture of the school mascot on the tokens (11.0 by 5.6 cm), and token economy data sheets.

The researchers also used the Behavior Rating Inventory of Executive Functioning 2 (BRIEF2) Teacher Form as a formal measure of self-regulatory functions (Gioia et al., 2015). Past studies showed that poorer executive functioning skills related to less effective delay of gratification strategies such as more temptation-focused behaviors (Jahromi et al., 2019). The BRIEF2 is an assessment tool designed to be completed by an adult, typically a teacher, who has
had extended contact with a child between the ages of 5 and 18 years old. The assessment consists of nine clinical scales (i.e., inhibit, self-monitor, shift, emotional control, initiate, working memory, plan/organize, organization of materials, and monitor) and two validity scales. Using a 3-point Likert-type scale ranging from “Never” to “Often”, teachers reported on how often specific behaviors have been a problem for the student (e.g., “Does not plan ahead for school assignments”). In this study, the researcher used a composite global score.

Psychometrically, the BRIEF2 has been reported to have high internal consistency ($\alpha_s = .80-.98$) and test-retest reliability ($r_s = .88$ for teachers). The assessment uses normative data with standardization sample ($N = 3,603$ total cases) matched by gender, age, ethnicity, and parent education level to U.S. Census statistics (Gioia et al., 2015). In the present study, Cronbach’s alpha for the BRIEF2 was highly reliable ($\alpha = 0.94$).

For the BRIC Task (delay discounting measure), the researcher used PowerPoint slides and a data sheet (Appendix A; from Critchfield & Atteberry, 2003 and Reed & Martens, 2011). Each PowerPoint slide consisted of a written direction, “Would you rather have,” on the top of the screen and had two boxes (7.62 cm X 7.62 cm) underneath it with the two reward values in each box. Each box represented a SSR value and a LLR monetary value. The researchers used the material and data sheet for only 1-month delay due to school closing as a response to the COVID-19 pandemic.

### 3.2.1.3 Procedure

#### 3.2.1.3.1 Preference Assessment

The researcher used a multiple stimulus without replacement (DeLeon & Iwata, 1996) to identify highly preferred backup reinforcers for the token economy. The researcher rank ordered individual preferences, with the first rank being the most preferred. Once the researcher collected
the preference-rank data for each participant, rank orders were averaged across participants. The researcher placed the most preferred item in the box with the highest value and the least preferred item in the box with the lowest value. There were seven boxes, wherein each box represented different costs (i.e., 1, 3, 5, 10, 20, 50, and 100 tokens).

3.2.1.3.2 Token Economy

The researcher conducted the study in a second-grade inclusion classroom using a classwide token economy system. A token economy system was used because tokens can function as an alternative approach to hypothetical choice or money (Reed et al., 2013; Staubitz et al., 2018) and are considered advantageous in that they bridge the delay between behavior and reinforcement (see Hackenberg, 2018)—an important concept in a delay discounting approach to treatment. A recent study that investigated delay discounting with a younger population suggested the need for further investigation in the use of alternative approaches or adaptations for children who may have difficulty understanding hypothetical constructs such as money or delay (Staubitz et al., 2018). Students in classrooms often earn reinforcers and backup reinforcers based on completion or performance—completion means when students earn tokens upon completing a task without consideration for accuracy (e.g., finishing a worksheet) while performance means when students earn tokens depending on completion and accuracy (e.g., completing a worksheet with 100% accuracy). Therefore, the researcher evaluated two different token production schedules, which is the rate at which delivery of tokens are made respective to the number of target behaviors emitted (Hackenberg, 2009; Kazdin & Bootzin, 1972), to evaluate the effects of the different earning schedules on students’ saving and spending behavior. The researcher termed the token production schedule “fixed” for completion because students earned
a fixed number of tokens each day while “variable” for performance because the number of tokens earned depended on learning performance that day.

3.2.1.3.2.1 Fixed Token Production Schedule

In the fixed token production schedule, each participant received two criterion cash per school day. Each participant received one criterion cash for reading and one for math at the end of each period contingent on completion of all activities for each period. Reading and math periods each lasted approximately 80 min every school day. At the end of each school day 20 min before dismissal time, the researcher called each participant to the prize store contingent on daily assignment completion. Then, the researcher asked, “How many criterion cash do you have today?” After the participant responded with the total number of criterion cash he or she possessed (i.e., remaining criterion cash from previous day plus those earned in the respective days), the researcher let the participant save or spend their criterion cash. If the participant decided to spend, he or she was allowed to select an item in accordance with the price of the item and the tokens the individual had earned. In other words, if the participant had seven criterion cash, he or she had the choice to go to 1, 3, or 5 prize box. Each participant was allowed to purchase one item per day at the end of the day. If the participant decided to save, he or she was allowed to keep the criterion cash earned on the respective day and previous days without exchanging the tokens for a prize. For each participant, the researcher recorded the total number of tokens and the number of tokens spent. The fixed token production schedule was in place for approximately 4 months during the fall semester (September to December).

3.2.1.3.2.2 Variable Token Production Schedule

The variable token production schedule procedure was identical to the fixed token production schedule, except that each participant received a variable number of tokens each day
depending on the number of learning objectives met. The teachers scripted out the learning objectives for each subject (i.e., reading, writing, and math) prior to the study and modified them depending on each participant’s academic needs. For example, a participant could earn two tokens by demonstrating mastery of adding 2-digit numbers using a number line and expanded form strategy. Each participant received a minimum of 0 tokens and maximum of 10 per day.

The variable token production schedule was in place for approximately 3 months during the spring semester (January to mid-March when the school closed due to the COVID-19 lockdown).

3.2.1.3.3 **Behavior Rating Inventory of Executive Functioning 2 (BRIEF2) Assessment**

Former teachers of each participant completed the BRIEF2 Teacher Form. The former teachers have had at least 6 months of time spent together during the prior school year. The researcher added the scores for a global score. These scores were translated into normative T scores, which provided information about an individual’s scores relative to the scores to other respondents in the standardization sample (Gioia et al., 2015). Higher global score reflects more problems with self-regulation (Gioia et al., 2015).

3.2.1.3.4 **Brief Intertemporal Choice Task (BRIC Task)**

The researcher administered the BRIC Task to obtain the indifference point at 1-month delay from the discounting task. The researcher used hypothetical, monetary choices of SSR and LLR and asked each participant a series of questions with binary choices. For each question, the participant was required to choose one of two monetary choices that had a contrasting length of delay and size of reinforcer. The researcher manipulated the amount of SSR values while holding the 1-month LLR delay value constant to determine the points at which each participant changed his or her choice from the LLR to the SSR. There were 22 possible indifference points (Appendix A; from Critchfield & Atteberry, 2003 and Reed & Martens, 2011).
All participants started with a binary choice between SSR of $50 and LLR of $100. The researcher provided the vocal antecedent, “Would you rather have $50 right now or $100 in one month?” If the participant chose the $50 available now, the next trial consisted of a smaller amount of money available now against $100 available in one month. However, if the participant chose $100 in one month, the next trial consisted of a larger immediate monetary amount and the delayed $100. The researcher collected data on a data sheet with all possible choices (Appendix A; from Critchfield & Atteberry, 2003 and Reed & Martens, 2011). Then, the researcher derived the subjective value of the $100 (i.e., indifference point) in terms of smaller amount of money available immediately (see Critchfield & Atteberry, 2003; Reed & Martens, 2011).

3.2.1.3.5 Data Analysis

3.2.1.3.5.1 Token Economy Data Analysis

The researcher first calculated the mean balance and mean peak point. The mean balance represented an overall number of tokens a participant kept in his/her account reflecting the daily changes in savings or expense, and the mean peak point represented an overall maximum number of tokens a participant accumulated until he/she decided to trade in. The mean balance was calculated by adding the total number of tokens each participant had each day and dividing the sum by the total number of trade-in days. The mean peak point was calculated by adding the “peaks” of each phase and dividing the sum by the total number of trade-in days. The researcher defined a “peak point” as a number that had a smaller preceding and following number in the data string. For example, if a data set showed 2, 4, 6, 8, 4, 6, 8, 10, 2, 4, 6, 2, with each number representing the balance of each day, then the data would have three peaks: 8, 10, 6, and a mean peak point of 8 (range, 6 to 10).

3.2.1.3.5.2 Descriptive Analysis
The researcher created a line graph displaying individual students’ mean peak point and mean balance in a fixed and variable token production schedule. The line graph displayed individual participants’ changes in the mean peak point and mean balance when the token production schedule changed from fixed to variable.

3.2.1.3.5.3 Statistical Analysis

Given the skewed distribution (i.e., variance was not normally distributed on a bell curve) of the indifference point at 1-month delay and BRIEF2 scores, the researcher conducted a nonparametric Spearman’s rho correlations ($r_s$) to determine whether students with higher mean peak point and mean balance have (a) higher indifference point at 1-month delay and (b) lower BRIEF2 global score.

3.2.1.4 Interobserver Agreement

A second observer independently collected data for the purpose of assessing interobserver agreement (IOA) of the delay discounting task. The researchers calculated trial-by-trial IOA by adding the number of binary choices in agreements, dividing by the total number of binary choices of agreements and disagreements and multiplying by 100%. The researchers obtained IOA data for 15 out of 16 BRIC tasks (94%) with 100% agreement.

3.2.3 Results and Discussion

3.2.3.1 Analysis of BRIEF2 and the BRIC Task

Table 2 shows the Spearman’s rho correlation among the mean peak point, mean balance, GEC scores, and the indifference point at 1-month delay obtained through the BRIC Task. In the fixed token production schedule, there was a significant positive correlation between the indifference point and the mean balance ($r_s = .53, p = .04$) but marginally significant with the mean peak point ($r_s = .49, p = .05$). In the variable token production schedule, there was a
significant positive correlation between the indifference point and the mean balance ($r_s = .56, p = .02$) and the mean peak point ($r_s = .59, p = .02$). Overall, students with higher indifference points were more likely to wait to exchange tokens (i.e., higher mean peak point and mean balance) in both fixed and variable token production schedules. Across both mean peak point and mean balance, the correlation was stronger in the variable token production schedule compared to the fixed schedule. In other words, students who discounted less saved and accrued more, and children’s performance on the BRIC Task had a strong relation with variable token production performances.

Students with higher GEC scores (i.e., reported to have more frequent self-regulation problems) were more likely to have a lower mean peak point and mean balance in both fixed and variable token production schedules. In the fixed token production schedule, there was a significant negative correlation between the GEC score and the mean balance ($r_s = -.59, p = .01$) as well as the mean peak point ($r_s = -.73, p = .001$). In a variable token production schedule, there was a significant negative correlation between the GEC score and the mean balance ($r_s = -.56, p = .02$) as well as the mean peak point ($r_s = -.51, p = .046$). Across both mean peak point and mean balance, the correlation was stronger in the variable token production schedule compared to a fixed schedule. Interestingly, there was not a significant correlation between the indifference point found by the BRIC Task and GEC score ($r_s = -.27, p = .31$), highlighting that the BRIC Task may capture elements of the delay context unique from other self-regulatory measures. Overall, students with higher GEC scores were more likely to have lower mean peak point and mean balance in both fixed and variable token production schedules. Across both mean peak point and mean balance, the correlation was stronger in the variable token production schedule condition compared to a fixed schedule. In other words, students who have higher self-
regulation save and accrue more tokens, and performance on the BRIC Task was highly correlated with variable token production performances. Collectively, higher executive functioning (i.e., better self-regulation) and higher indifference point at 1-month delay maps onto saving more.

3.2.3.2 Mean Peak Point and Mean Balance

Figure 1 shows descriptive data on the change of individual participants’ mean balance and peak point when the token production schedule changed from fixed to variable. All students, except for one in the mean peak point and two for mean balance, showed an increase in mean peak point and mean balance when the token production schedule changed from fixed to variable. In a fixed token production schedule, the classwide mean peak point was 12.47 (range, 3.57 to 78.5) and classwide mean balance was 9.77 (range, 2.93 to 51.12). In a variable token production schedule, the classwide mean peak point was 19.20 (range, 6 to 64) and classwide mean balance was 15.70 (range, 4.11 to 33.77).

Figure 2, 3, 4, and 5 display the mean peak point and mean balance in a fixed and variable token production schedule between groups of students with and without disabilities, students who do and do not receive free/reduced lunch (i.e., a proxy measure of a participant’s socio-economic status), students who are and are not ELL, and male and female students, respectively. From a descriptive standpoint, compared to students with disabilities (Figure 2, top panel), students without disabilities (Figure 2, bottom panel) generally exhibited greater sensitivity to the token production schedule change. In other words, the mean peak point and mean balance data showed a greater change when the token production schedule changed from fixed to variable for students without disabilities, while they showed minimal change for students with disabilities.
This general pattern of descriptive data was also found for students who received regular lunch (Figure 3) and those who were not ELLs (Figure 4). That is, the mean peak point and mean balance data generally showed a greater change from fixed to variable production schedule for students in these groups relative to their counterpart peers.

Compared to the female students’ data (Figure 5, top panel), male students’ data (Figure 5, bottom panel) are more distributed for mean peak point and mean balance in both fixed and variable token production schedules. The data points were more concentrated for female students across both token production schedules.

Overall, Study 2 showed that the BRIC Task was a unique evaluation and predicts actual student behavior in token economy systems. Some major findings include: (1) BRIC Task predicts classroom behavior, (2) students who discounted less and had higher self-regulation saved and accrued more tokens, (3) a variable token production schedule better represented saving and spending behavior compared to a fixed token production schedule, (4) BRIC Task may capture a different element of the delay context than what is tapped by the BRIEF2, which captures a broader range of children’s self-regulation behaviors, and (5) students with disabilities, those who received free/reduced, and who were ELL generally exhibited less sensitivity to the token production change compared to their counter-groups.

3.3 General Discussion

Delay of rewards is present in many settings including classrooms with younger children (Reed & Martens, 2011). This warrants the need for more translational studies on the application of delay discounting for younger population in applied settings (Staubitz et al., 2018). However, the complex nature of deriving meaningful numbers impedes its widespread use among educators and clinicians (Critchfield & Reed, 2009). The present study attempted to fill the gap
in literature pointed by Staubitz et al. (2018) and introduce a more accessible, user-friendly method for educators and clinicians to measure sensitivity to reward delays for school-aged children. While other discounting procedures for children show promise in research settings (e.g., Miller, 2019), the current approach leverages a larger body of existing research on the protocol and does not require specific technologies typically afforded to scientists (e.g., survey software).

First, the findings from the two studies demonstrated that the delay discounting task at 1-month delay is reliably associated with students’ classroom behavior. This means educators and clinicians can easily quantify sensitivity to reward delays without involving complex mathematical equations and knowledge in statistics and software (Motulsky & Christopoulis, 2006). This is not to replace the existing methods to measure delay discounting, but rather add an alternative measure that individuals can use without deriving additional data. The BRIC Task can thereby serve as a valid, alternative tool to measure sensitivity to reward delays in applied settings for younger populations.

Second, students who discounted less and who were a) reported by teachers to have higher self-regulation, and b) saved and accrued more tokens. These findings suggest that hypothetical delay discounting assessments and those from mainstream psychology that capture self-regulation more broadly do show associations with meaningful behavior in applied settings. Critchfield and Kollins (2001) posed a question on whether hypothetical delay discounting assessments translate to or predict observable behaviors reinforced with real rewards. Similar to past studies, (Kirby & Marakovic, 1996; Reed & Martens, 2011; Richards et al., 1999), our findings showed an adequate level of correlation between hypothetical delay discounting task at 1-month delay for children and their responsiveness to delayed classroom rewards in two
different token production schedules. Further, to the best of our knowledge, there has not been research that directly tested associations between school-aged children’s spending and saving behavior in token economies and discounting rate and teachers’ reports of self-regulation. Token economies are widely used across educational and clinical settings and have been identified as an effective evidence-based classroom management strategy (Simonsen et al., 2008). This means that the possible application of the findings is extensive.

Third, the pattern of descriptive results showed individual changes in response to token production schedule change. Students with disabilities, students who receive free/reduced, and students who are ELL generally exhibited less sensitivity to the token production change compared to their counter-groups. These data were only descriptive. Future studies may want to examine whether factors like social competition or student motivation is related to students’ sensitivity to rewards. Further, individual students showed higher levels of saving under a variable token production schedule compared to a fixed one. Students may have been more motivated to save when they had the opportunity to earn tokens based on their performance.

3.3.1 Implications for Future Research

There are several implications for future research. First, the present study suggests the role of the token production schedule or, in other words, response requirements in the preference of SSR or LLR. Past studies showed that preference for the LLR varies across populations and experimental conditions. Some researchers argued that organisms tend to value rewards available sooner rather than later (Hackenberg & Vaidya, 2003; Hyten et al., 1994) and others have argued that students tend to prefer LLR over SSR (DeLeon et al., 2014; Fienup et al., 2011; Ward-Horner et al., 2014). However, some of these studies did not directly manipulate the token production schedule or were not implemented in an applied setting (Hackenberg & Vaidya,
2003; Hyten et al., 1994), and some studies that investigated the response requirements did not use a token economy system (Fienup et al., 2011; Ward-Horner et al., 2014). Thus, future studies should investigate how manipulating the token production schedule in an applied setting affects individuals’ preference for SSR or LLR.

A second implication is the development of an intervention that promotes optimal choice making or behavior consistent with high delay of gratification in educational settings. Previously, researchers have taught individual optimal choice making by immediately delivering large magnitude reinforcers and then progressively increasing the delay across students with developmental disabilities (Vollmer et al., 1999), emotional and behavioral disorders (Staubitz et al., 2020), brain injury (Dixon, & Tibbetts, 2009), and ADHD (Binder et al., 2000). Research also showed that perceived importance of peer and parental attitudes toward saving money help explain children’s saving behavior (Te’eni-Harari, 2016). Given the possible role of peer influence on saving behavior and possibility of teaching optimal choice making behavior, adding a social competition element to this procedure may facilitate the process of improving delayed gratification for individuals with disabilities. For future studies, it may be worth investigating the effects of putting students in teams or competing against the teacher along with the fading procedure.

3.3.2 Implications for Educators

The findings also have implications for educators. Educators should use the BRIC Task to gather information on a client or student’s sensitivity to delayed backup reinforcers or general level of self-regulation. Educators working in classrooms with younger populations can use the BRIC Task as a quick measure for individuals’ sensitivity to reward delays or self-regulation. An individual whose indifference point is larger is less sensitive to reward delays. In other words,
the larger the indifference point is, the more likely an individual saves tokens for a backup reinforcer and has greater self-regulation. With such information, educators can arrange effective reward systems. If an individual scores higher in the BRIC Task (i.e., higher indifference point), educators could use leaner token production rate which requires greater number of responses to earn a reinforcer. If a student scores lower (i.e., lower indifference point), educators can utilize richer token production rate which requires smaller number of responses to earn a reinforcer and consider using delay fading of LLR to build delay of gratification. Educators can systematically fade the number of responses required to earn a reinforcer (i.e., progressively increase the number of responses) to facilitate delay of gratification. These individuals will also be more likely to spend the tokens warranting the need for more frequent token-earn schedule.

Further, a variable token-production schedule (i.e., a schedule that allows individuals to earn tokens on based on performance) is recommended for greater saving behavior. Educators should have clients or students earn tokens based on performance rather than completion to encourage saving token behavior. This way, educators can increase token saving, ensure that the client or student comes in contact with the LLR, and potentially enhance delay of gratification in a naturalistic setting. Thus, the saving behavior will come under the control of natural contingencies, facilitating generalization.

3.4 Limitations

The study is not without limitations. In the classwide token economy procedure, there were limited types of rewards in the prize store. There were 16 types of rewards in total, which resulted in two or three types of items in each prize box. The researcher addressed this limitation by diversifying the colors, shapes, sizes, patterns, and materials of the same item. For example, a yellow pencil used in the preference assessment would represent different colored and patterned
pencils. The researcher also rotated the different types of items every two to three days to ensure that the students were not satiated. Moreover, given the natural classroom setting, school holidays and events were inevitable. The class did not have trade-in sessions on these days, and the researchers did not count these days for analysis to keep the data consistent across all students.

Further, there were limitations to the implementation of the assessments in Study 2. Initially, the researcher planned to conduct a full hypothetical delay discounting assessment with children but was unable to complete due to school closure as a response to the COVID-19 pandemic. The researchers serendipitously discovered that the 1-month delay can be sufficient enough to predict actual student behavior and inform individuals about school-aged children’s sensitivity to reward delays. While the serendipitous findings are noteworthy, the study would have benefited from doing a full delay discounting assessment. The researchers would have been able to compare the 1-month delay to other delays, further supporting the use of 1-month delay as a proxy for a full delay discounting assessment. Future studies should replicate the procedure with the complete assessment with eight indifference points to compare the delays using the same participant pool.

Additionally, in Study 2, the researchers used an average preference instead of individual preferences to identify backup reinforcers preference assessment. Some participants’ lowest-ranked items had a higher cost while some highest-ranked items were lower cost – when examining cost at an individual level. This means that there may have been limited reinforcing value of the backup reinforcers, which means that saving could have been a product of not finding any items valuable. However, 13 out of 17 participants (76%) selected the top-ranked reinforcer as their top-three choices, which means that the top-ranked reinforcer would have had
the reinforcing effect for the majority of participants. The remaining four participants (24%) ranked the top-ranked reinforcer as 4th, 6th, 8th, and 8th out of 16 items. The lowest-ranked reinforcer, on the other hand, was selected as the bottom-three choices by only four out of 17 participants (24%), limiting the value of the item associated with low cost. While the top-ranked reinforcer had reinforcing value for most participants, the discrepancy between individual rank ordering and item price should be taken into consideration when interpreting the results. Future studies should also incorporate reinforcer assessments to ensure that preferred items functioned as reinforcers.

Another limitation is the lack of treatment fidelity data. Treatment fidelity ensures that the treatment was implemented as planned and eliminates threats to bias (Reichow et al., 2018). Future studies should incorporate a checklist for the token economy and BRIC Task procedures to ensure high levels of treatment fidelity. Also, addition of social validity measure that evaluates teacher preference for the BRIC Task over simple observations would have supplemented the findings regarding the utility of the BRIC Task in educational settings. This way, all consumers of the program contribute to the evaluation of acceptability and viability of the program goals, methods, and outcomes (Schwartz & Baer, 1991).

Lastly, implementation of fixed and variable schedule was not systematically manipulated, which may limit the comparison of the two token economy systems. Given the educational setting, the researcher implemented the same contingencies across all students in the same class as part of a classwide token economy system. However, because the variable token production schedule condition was introduced a few months later, participants may have matured and engaged in more saving behavior since older children save more money compared to younger children (Mischel & Mischel, 1983). To address this problem, future studies should
systematically manipulate the system to determine the effects of the different token production schedule on students’ saving and spending behavior.

3.5 Conclusion

Despite the limitations, the two studies demonstrated that the indifference point at 1-month delay can adequately predict actual behavior and introduced a more accessible, user-friendly tool for educators and clinicians to measure sensitivity to reward delays for younger population. The study extends previous research by (a) establishing a translational utility of delay discounting measures to a naturalistic setting, (b) developing an alternative tool that warrants a widened use of delay discounting measures across fields, (c) experimentally demonstrating that children who discount less and have more self-regulation save and accrue more tokens, and (d) comparing the effects of different token production schedule on different groups. The findings may serve as the foundation for the possible future development of an intervention promoting optimal choice for younger population who are at higher risk for poor long-term outcomes.
References


Table 1

Summary of Reanalysis of Reed & Martens (2011)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median 1-month indifference point</td>
<td>$30 (IQR = 7.00 to 70.00)</td>
<td>$50.00 (IQR = 6.25 to 80.00)</td>
</tr>
<tr>
<td>Median AUC</td>
<td>0.13 (IQR = 0.04 to 0.24)</td>
<td>0.12 (IQR = 0.02 to 0.29)</td>
</tr>
<tr>
<td>Correlations between the 1-month indifference point and AUC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$r_s = .45***$</td>
<td>$r_s = .58***$</td>
</tr>
<tr>
<td>Correlations between the 1-month indifference point and the difference scores between immediate and delayed classroom contingencies</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$r_s = -0.33*$</td>
<td></td>
</tr>
<tr>
<td>Test-retest reliability of the 1-month indifference point</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$r_s = 0.80***$</td>
<td></td>
</tr>
<tr>
<td>ED50</td>
<td>34.27 days (SD = 40.14)</td>
<td></td>
</tr>
</tbody>
</table>

Note. AUC = Area Under the Curve; ED = Effective Delay.

*p < .05, *** p < .001
Table 2

Correlations among Mean Peak Point, Mean Balance, BRIEF2, and Subjective Value of $100 at 1-Month Delay

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mean B (Var)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Mean PP (Var)</td>
<td>.80**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Mean B (Fixed)</td>
<td></td>
<td>.55*</td>
<td>.73**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Mean PP (Fixed)</td>
<td>.67**</td>
<td>.76**</td>
<td>.96**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. BRIEF2 GEC</td>
<td>-.56*</td>
<td>-.51*</td>
<td>-.59*</td>
<td>-.73**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Subjective Value of $100 at</td>
<td>.56*</td>
<td>.59*</td>
<td>.53*</td>
<td>0.49</td>
<td>-0.27</td>
<td></td>
</tr>
</tbody>
</table>

1-Month Delay

Note. B = Balance; PP = Peak Point; Var = Variable Token Production Schedule; Fixed = Fixed Token Production Schedule; BRIEF2 GEC = Global Executive Composite score.

* p < .05, ** p < .01
Figure 1

Individual Change of Mean Peak Point and Mean Balance When Token Production Schedule Changes

Note. This figure shows the individual participant’s change of mean balance and mean peak point when the token production schedule changes from fixed to variable.
Mean Peak Point and Mean Balance between Students with and without Disabilities

Note. The scatterplots compare the mean peak points (left column) and mean balance (right column) between students with and without disabilities in fixed and variable token production schedule. Each panel represents the change in one group’s level of mean peak point and mean balance in response to token production schedule change from fixed to variable.
Figure 3

Mean Peak Point and Mean Balance for Students who Receive Free/Reduced Lunch and Regular Lunch

Note. The scatterplots compare the mean peak points (left column) and mean balance (right column) in fixed and variable token production schedule between students who receive and do not receive free/reduced lunch. Each panel represents the change in one group’s level of mean peak point and mean balance in response to token production schedule change from fixed to variable.
Figure 4

Mean Peak Point and Mean Balance between Students who are ELL and Not

Note. The scatterplots compare the mean peak points (left column) and mean balance (right column) in fixed and variable token production schedule between students who are and are not English Language Learners (ELL). Each panel represents the change in one group’s level of mean peak point and mean balance in response to token production schedule change from fixed to variable.
Figure 5

Mean Peak Point and Mean Balance for Female and Male Students

Note. The scatterplots compare the mean peak points (left column) and mean balance (right column) in fixed and variable token production schedule between male and female students. Each panel represents the change in one group’s level of mean peak point and mean balance in response to token production schedule change from fixed to variable.
Appendix

Example of the BRIC Task Datasheet (from Critchfield & Atteberry, 2003; Reed & Martens, 2011)

Would you rather have _____ now or _____ in _____?

In each box, blacken in the circle beside the money option that the participant chooses. Then follow the line that leads from their answer to the next question. When their answer leads to no more questions, go on to the next page.

START HERE

$500 paid right now

- OR -

$1000 paid in 1 month

$8 paid right now

- OR -

$1000 paid in 1 month

$25 paid right now

- OR -

$1000 paid in 1 month

$55 paid right now

- OR -

$1000 paid in 1 month

$150 paid right now

- OR -

$1000 paid in 1 month

$250 paid right now

- OR -

$1000 paid in 1 month

$350 paid right now

- OR -

$1000 paid in 1 month

$450 paid right now

- OR -

$1000 paid in 1 month

$650 paid right now

- OR -

$1000 paid in 1 month

$750 paid right now

- OR -

$1000 paid in 1 month

$850 paid right now

- OR -

$1000 paid in 1 month

$950 paid right now

- OR -

$1000 paid in 1 month

$950 paid right now

- OR -

$1000 paid in 1 month

Note. Republished with permission from John Wiley & Sons.
Chapter 4: A Preliminary Evaluation of an Online Classwide Progressive Delay Training on Advantageous Choice-Making Behavior in General Education

Outline

Advantageous choice-making behavior is the behavior of selecting a larger, delayed outcome over a smaller, immediately available reinforcer. The current study investigated the effects of a classwide progressive delay training procedure on students’ advantageous choice-making behavior given that no studies have investigated the effects of a progressive delay training 1) for larger groups of children or 2) in general educational settings. The intervention used progressive delay, which refers to incrementally increasing the temporal delay for a more delayed, advantageous outcome, and choice-making opportunities to promote second grade students’ choice of a larger reinforcer associated with a delay. The results showed mixed results – the intervention was effective at increasing advantageous choice-making behavior for students performing on and below grade-level for math, but the effects were not shown in students performing above grade-level. Future applications in educational settings and research are discussed.

Keywords: progressive delay training, delay of gratification, general education, delay discounting, self-control.
Online Classwide Progressive Delay Training on Advantageous Choice-Making Behavior in General Education

Delay of gratification is a subconstruct of self-regulation and is defined as the behavior of selecting a more advantageous, delayed outcome over a smaller, immediately available reinforcer (Jahromi, 2017). A series of studies, including the famous marshmallow experiment (Mischel et al., 1988), showed that a child’s delay of gratification relates to later cognitive and social development and even brain structure (Casey et al., 2011; Mischel et al, 2010). Studies also showed that children who wait longer are more academically competent, verbally fluent, attentive, rational, planful, and able to handle stress better as adolescents (Mischel et al., 1988). Children with poor self-regulation skills make less academic progress (McClelland et al., 2007) and tend to experience more anxiety, depression, and aggressive behavior problems (Martel et al., 2007). In the long run, children with poor delay of gratification are at higher risk for poor health outcomes including obesity and drug dependency (Moffitt et al., 2011). Further, having strong self-regulation skills was related to school success in both math and reading (Adams & Snowling, 2001; Connor et al., 2010; Day et al., 2015; St. Clair-Thompson & Gathercole, 2006).

A behavior analytic concept that represents advantageous choice-making behavior is delay discounting, which is defined as devaluation of a reinforcer as a function of temporal delays. It is a concept within behavioral economics whereby researchers study an organism’s preference between smaller, sooner reinforcers (SSR) and larger, later reinforcers (LLR) and quantifies an organism’s sensitivity to reinforcers across temporal dimensions. Similar to studies on delay of gratification, past studies on delay discounting showed that myopic and immediate reinforcer preferences related to many maladaptive behaviors including cigarette smoking (Mitchell, 1999), drug and alcohol abuse (Richards et al., 1999), overeating (Weller et al., 2008),
and health problems (Daugherty & Brase, 2010). In fact, younger populations discount the value of delayed reinforcers as adults do. Reed and Martens (2011) examined delay discounting of sixth-grade students in a classroom setting and compared the effects of varying reinforcement contingencies on on-task behavior. The researchers found that LLR were less effective in improving on-task behavior compared to immediate SSR for students with higher discounting rate (i.e., students for whom delays to reinforcement devalue the stimulus at a faster rate). Together, the studies suggest that delay discounting may be a promising approach to understanding younger individuals’ advantageous choice-making behavior.

Although delay of gratification and delay discounting are similar constructs, delay of gratification is a broader concept that explains advantageous choice-making behavior while delay discounting is an economic valuation process that informs the child’s delay of gratification. Another difference lies in how each construct is measured (Göllner et al., 2018). Delay of gratification quantifies behavior directly while delay discounting often involves hypothetical choices. Delay of gratification assessment with children involves presenting an immediate, smaller reinforcer (e.g., one marshmallow or cookie) and a delayed, larger reinforcer (e.g., two marshmallows or cookies). Delay of gratification is measured using the time the child can wait for the larger reinforcer (Mischel, 1974). Delay discounting, on the other hand, involves a series of question with binary intertemporal choices using varying monetary values and delays (Kirby, 1999). For example, a question would ask if the participant would prefer $1 right now or $10 tomorrow. Typically, researchers would find several indifference points, which are points at which the value switches from the LLR to the SSR (Reed et al., 2013) and use them to find discounting parameters (e.g., $k$ values; see Odum, 2011) through nonlinear modeling (see Gilroy et al., 2017) and area under the discounting curve (AUC; e.g., Myerson et al., 2001). Past studies
also showed that delay of gratification and delay discounting were statistically unrelated for children ages 7- to 9-year-olds (Burns et al., 2020); however, as Burns et al., (2020) suggested, it should be noted that the methodological differences between delay discounting and delay of gratification measures make it difficult to be confident in the finding yet. Nonetheless, both concepts operationalize an organism’s behavior of selecting a more delayed, advantageous outcome over a smaller, immediately available reinforcers and highlight the importance of advantageous choice-making behavior in younger population.

While genes play a role in the development of children’s advantageous choice-making behavior (Reif et al., 2009), the environment (e.g., school context and teachers) also affects how children make advantageous choices. Children are constantly contacting new stimuli in the environment that then provide the experiential stimuli for contacting other experiences (Greer et al., 2017). It is important that children contact stimuli for positive behavior change especially in settings they spend a substantial amount of time such as schools. In the United States, children spend an average of 6.64 hr in the school each day and spend 180 days in a year in school (United States Department of Education, 2007-2008). This means that educators spend a considerable amount of time with students and are responsible for creating an environment that supports socially significant behavior such as advantageous choice-making behavior.

Behavior scientists have developed a method to teach children advantageous choice-making behavior that employs a progressive delay schedule of reinforcement. Past studies that have successfully taught individuals to select the LLR by progressively increasing the temporal delay (these studies called this procedure a “self-control training”; Dixon & Falcomata, 2004; Dixon & Tibbetts, 2009). Progressive delay training has been used in isolation (Dixon & Falcomata, 2004; Schweitzer & Sulzer-Azaroff, 1988) and in combination with rules (Benedick
& Dixon, 2009), intervening activities (Dixon et al., 2003), and signaled delays (Vollmer et al., 1999). Another variation of a progressive delay training is incorporating choice opportunities (Dixon & Tibbetts, 2009). The procedure was based on the finding that humans prefer greater control over choice-making opportunities (Fisher et al., 1997; Tiger et al., 2006). In the progressive delay training procedure involving choices, an individual was given three choices including doing nothing and earning the SSR, engaging in the maximum required number or duration of the target behavior to earn the LLR, and engaging in the target behavior with a progressively increasing temporal delay to earn the LLR. When the participants selected the progressively delayed choice associated with the LLR, the experimenters incrementally increased the temporal delay for the LLR until the participants reached the maximum delay (Dixon & Tibbetts, 2009; Dixon & Falcomata, 2004). The progressive delay training involving choices successfully changed individuals’ preference to LLR.

Nevertheless, a more recent study showed mixed results on the effectiveness of a progressive delay training procedure to teach students advantageous choice-making behavior in a school setting (Staubitz et al., 2020). The researchers implemented a modified progressive delay training procedure with six participants enrolled in a public, self-contained school for students with emotional and behavioral disorder. While the traditional progressive training procedure alone did not improve rule-following behavior, a modified version with a rationale and rule for selecting delayed reinforcer effectively increased rule-following behavior for half of the participants. The researchers suggested that the training procedure may not be feasible for use in a pull-out setting and recommended an application of the procedure in a general education setting (Staubitz et al., 2020). It would be beneficial for educators to utilize an intervention to teach advantageous choice-making behavior in an educational setting because educators often
evaluate students’ level of self-regulation and indicate it on report cards as an area of behavior or work habits. Although delay of gratification is educationally and socially important, existing literature only explored the progressive delay training on a one-to-one basis where the experimenter implemented the procedure with one participant at a time. However, there is often limited resource such as time and labor (e.g., teacher and assistants) in a general education setting, which highlights the need for a more efficient training procedure such as a classwide intervention.

The purpose of the current study was to determine the effects of a classwide progressive delay training with choice opportunities on students’ advantageous choice-making behavior. To date, studies that have used the progressive delay or self-control training procedure had a small number of participants and have been conducted in a tightly controlled environment on a one-to-one basis. To the best of the researcher’s knowledge, no studies have investigated the effects of the delay of gratification training on a group-basis for children in a general educational setting. During the training, this study specifically used addition and subtraction math fact fluency to incorporate a second-grade curricular objective that aligned with the common core standard. Math fact fluency was used for two main reasons. First, fluency is a combination of fast and accurate responding which is critical to student success (Haring & Eaton, 1978). Students who can fluently complete math facts are more fluent in completing more advanced math tasks (Skinner, 1996), highlighting the importance of teaching math fact fluency. Second, studies showed that students with higher math performance had greater level of self-regulation (St. Clair-Thompson & Gathercole, 2006; Wang et al., 2019). This means that students’ level of math should be taken into consideration to determine the students’ level of advantageous choice-
making behavior prior to the intervention. Thus, the participants were grouped based on their math level to group students by their needs of the intervention.

The current study extended the past research by implementing the progressive delay training across three groups of second graders using addition and subtraction math fact fluency. The researchers aimed to teach more advantageous choice-making behavior in a general classroom setting by promoting their choice of a larger reinforcer associated with a progressively increasing temporal delay. The entire study was conducted online as part of a measure during the COVID-19 pandemic and was conducted during math period. Research questions the current study will address include: (a) will students choose the LLR more frequently after the implementation of an online classwide progressive delay training, and (b) will the subjective value of $100 at 1 month delay improve (i.e., have larger indifference point, indicating greater tolerance for delay) after the implementation of the delay of gratification training?

4.1 Method

4.1.1 Participants

Participants included 20 second-grade students (10 males) with a mean age of 7.63 years at the onset of the study ($SD = 0.30$ years; see Table 1 for detailed demographic information). Participants were selected from an online inclusion second grade classroom housed in Title 1 public elementary school located in a suburb outside of a large metropolitan area with preschool to second grade classrooms. The sample consisted of 6 children with disabilities ($M$ age = 7.65, $SD = 0.15$ years) who had an Individualized Education Program (IEP) and 14 children without disabilities ($M$ age = 7.63, $SD = 0.07$ years). The sample consisted of students who were White (35%), Black (20%), Hispanic (20%), multi-racial (15%), and Asian (10%). Four children in the sample were classified as English Language Learners (20%).
The participants were divided into three groups based on their math performance on the *i-Ready® K-12 Adaptive Math Diagnostic* (Curriculum Associates, LLC, 2017) assessment conducted in October (i.e., beginning of year). The participants were grouped based on their math levels because studies showed that academic success in math related to self-regulation (St. Clair-Thompson & Gathercole, 2006). Grouping students based on their math level allowed the researchers to group students by their needs of the intervention. In other words, students with lower math performance would have lower self-regulation and benefit more from the intervention by learning to select more larger, later reinforcers. Group 1 consisted of 7 participants (\(M_{\text{age}} = 7.64, SD = 0.14 \text{ years}\)) who scored on-grade level (\(M = 403.57, SD = 4.16\)), Group 2 consisted of 6 participants (\(M_{\text{age}} = 7.50, SD = 0.11\)) who scored below-grade level (\(M = 403.10, SD = 8.38\)), and Group 3 consisted of 7 participants (\(M_{\text{age}} = 7.74, SD = 0.09\)) who scored above-grade level (\(M = 443.14, SD = 12.51;\) see Table 1 for participant information in each group). One student’s data from the below-grade level group was omitted from the mean calculation because this participant received a high math score despite anecdotal evidence of low performance in math. This may have happened because i-Ready assessments were taken from home under the parents’ supervision. For this participant, we used another measure – initial assessment or Reflex Math Fact Fluency Program (Explore Learning, 2018), which is a computerized intervention used to improve math fact automaticity. This participant received 0 in the initial assessment, which meant that the student was fluent in zero one-digit addition and subtraction math facts. Thus, this participant was added to the below-grade level group. All 20 participants had the prerequisite technology skills required to join instruction using online platforms and complete online daily assignments. This included joining Google Meet, a
videoconferencing platform, sessions on time according to the daily schedule, turning the camera and microphone on and off, opening the assignment, completing it, and turning in the work.

4.1.2 Setting and Materials

The study took place in an online classroom setting where all instruction was provided using Google Meet. The classroom was housed in a Title 1 public elementary school located in a suburb outside of a large metropolitan area with pre-school to second grade classrooms. The classroom implemented evidenced based behavioral and academic interventions and implemented the Comprehensive Application for Behavior Analysis to Schooling (CABAS®, Accelerated Independent Learner program; Greer, 1994) model. This model included one head teacher and one teaching assistant who implemented a scientific approach to pedagogy, learning, curriculum, and classroom management, and all instruction was individualized to each student’s repertoire.

Materials included a computer, camera, Microsoft PowerPoint, Google Meet, and links to online activities. All participants received a Chromebook provided by the district. Chromebooks distributed by the school district had a camera and access to internet. All students and teachers were provided access to Google Meet using the district account. The researchers used Microsoft PowerPoint to create the choice baseline and delay of gratification training slides (see Appendix A and B for an example). Mystery prizes included links to online activities such as games, museums, zoos, and games with the teacher.

4.1.3 Dependent Variable

The dependent variables were the (1) mean number of math facts completed for each group, (2) percentage of LLR selection for each participant, and (3) subjective value of $100 obtained using the Brief Intertemporal Choice (BRIC) Task (Kim et al., under review). All data
were collected during the school year across five months from November to March. First, the researcher collected the number of math facts completed by each participant each day. This means that the number of attempts of solving each problem was recorded. The problems were math problems previously mastered. The researchers provided additional practice on math facts to increase fluency (i.e., accuracy and speed) since the common core standard states that “by the end of Grade 2, [students] know from memory all sums of two one-digit numbers” (CCSS.MATH.CONTENT.2.OA.B.2). To calculate the mean number of math facts completed for each group, the total number of completed math facts for all students in the group was divided by the number of students of the group.

Second, the researcher obtained the percentage of LLR selection for each participant to measure each participant’s degree of advantageous choice-making behavior. In the choice baseline phase (see Procedure), the number of SSR and LLR selection per participant was recorded. Each participant emitted a selection response by raising his/her hand for one of the binary choices. The researcher divided the total number of LLR selection by the total number of choice opportunities per participant, multiplying by 100.

Third, Brief Intertemporal Choice (BRIC) Task (Kim et al., under review) was used to measure an individual participant’s sensitivity to delayed reinforcer by determining the indifference point at 1-month delay. Indifference points inform researchers the value at which individuals switch from the LLR to SSR (Reed et al., 2013). The researcher used hypothetical, monetary choices of SSR and LLR and asked each participant a series of questions with binary choices. For each question, the participant was required to choose one of two monetary choices that had a contrasting length of delay and size of reinforcer. The researcher manipulated the amount of SSR values while holding the 1-month LLR delay value constant to determine the
points at which each participant changed his or her choice from the LLR to the SSR. There were 22 possible indifference points (Critchfield & Atteberry, 2003; Reed & Martens, 2011). All participants started with a binary choice between a SSR of $50 and a LLR of $100. The researcher provided the vocal antecedent, “Would you rather have $50 right now or $100 in one month?” If the participant chose the $50 available now, the next trial consisted of a smaller amount of money available now against $100 available in one month. However, if the participant chose $100 in one month, the next trial consisted of a larger immediate monetary amount and the delayed $100. The researcher collected data on a data sheet with all possible choices (from Critchfield & Atteberry, 2003 and Reed & Martens, 2011). Then, the researcher derived the subjective value of the $100 (i.e., indifference point) in terms of the smaller amount of money available immediately (Critchfield & Atteberry, 2003; Reed & Martens, 2011).

4.1.4 Procedure

4.1.4.1 Natural Baseline

Before and after the intervention and choice baseline phase, the researcher implemented the natural baseline phase. The natural baseline phase consisted of posting Morningside math fact fluency sheets (Johnson, 2008) with addition and subtraction facts within 20 on Google Classroom each day. Each sheet was a review sheet consisting of 100 math facts that participants had previously mastered. The researcher posted the worksheet using Google Slides with answer boxes available for each fact. The worksheet was posted on Google Classroom at the beginning of school at 9 a.m. and removed at the end of school at 2 p.m. The worksheet was only available during school hours to ensure that the natural baseline measured the participants’ preference of math over other competing reinforcers such as online and offline activities. The researcher provided a vocal antecedent in the beginning of the day that the worksheet was available and that
the math facts were optional. The worksheet was only visible for the assigned participants. At the end of the day, the worksheet was removed from Google Classroom, and the researcher counted the number of math facts completed by each participant. Then, the mean number of math facts completed for the group was calculated.

4.1.4.2 Choice Baseline

Once there was stable responding in natural baseline, the researcher introduced a choice baseline phase. At the end of math period approximately at 1 p.m., the researcher vocally asked the participant group to stay on the online platform for a game. Then, the researcher shared the screen to show a slide consisting of two boxes adjacent to each other. Each box represented a SSR or LLR choice (see Appendix A for example). The LLR choice was fixed because the number of math facts required to receive the delayed reinforcer remained unchanged at 50 facts. The researcher provided a vocal direction that the participants should close their eyes and raise their hand for one of the choices. The SSR box consisted of the written contingencies, “No more math! 2 minutes of mystery prize” while the LLR box consisted of “50 math facts, 10 minutes of mystery prize.” If the majority of the participant group voted for SSR (e.g., four or more out of seven participants in Group 1; four or more out of six participants in Group 2; three or more out of five participants in Group 3), the researcher did not provide any additional math facts and revealed the mystery prize along with the link to the online activity. The participants were allowed to engage in the online activity for 2 min. If the majority of the participant group voted for LLR, the researcher vocally presented 50 addition and subtraction math facts (e.g., the researcher saying, “what is 4+6?” along with other addition and subtraction math facts using one-digit numbers). A vocal antecedent was provided instead of a written format (e.g., worksheet) to ensure that the researchers (who were also the first and second authors) provided
immediate consequences following participant choices. If the researcher provided a worksheet once the participants voted for the LLR, the participants would have had to complete the worksheet independently and then return to the video meet to receive the reinforcer. It was possible that some participants would not have completed the math problems and not have returned to the video meet for the reinforcer since math was the last academic period each day. Thus, the researcher had the participants stay on to respond to math problems presented vocally. The participants emitted a vocal or written response within 5 s of the vocal antecedent. The researcher provided vocal praise for correct responses. No consequence was provided for incorrect responses. Instead, after 5 s of the vocal antecedent, the researcher vocally provided the correct response. Once the group completed 50 math facts, the researcher revealed the mystery prize along with the link to the online activity. The participants were allowed to engage in the online activity for 10 min. The researcher ensured that the participant stayed on the online prize link for the respective duration by monitoring student activity through Go Guardian\(^3\). The researcher closed the page once the participant engaged in the activity link for the respective amount of time.

4.1.4.3 Delay of Gratification Training

The delay of gratification training was introduced once there was stable responding in choice baseline. The training condition was identical to the choice baseline phase with an addition of a third choice. The third choice, progressive larger later reinforcer [P-LLR choice, involved 10 min of mystery prize contingent on the completion of a progressively increasing number of math facts. The researcher started with completing zero math facts and increased the

\(^3\) Go Guardian is an online platform that allows teachers to monitor students’ screens, open a web page using a URL, close a page, and chat or call students when needed.
number of math facts by five facts progressively until participants completed 50 math facts in a session. The number of math facts was increased when the majority of participants in a group voted for the P-LLR choice across two consecutive sessions. Once the group chose the P-LLR with 50 math facts across two consecutive sessions, the researcher returned to the choice baseline followed by the natural baseline phase.

4.1.5 Experimental Design

The researcher used a changing criterion design embedded within a concurrent multiple baseline across groups design (Horner et al., 2005; Hartman & Hall, 1976) to assess the effectiveness of the delay of gratification training on choice-making behavior. The order of treatment was: (A) pre-intervention natural baseline and BRIC Task assessment, (B) pre-intervention choice baseline, (C) delay of gratification training, (D) post-intervention choice baseline assessment, and (E) post-intervention natural baseline and BRIC Task assessment. The researcher collected natural and choice baseline data until each group showed steady state responding with little variation. Then, the researcher implemented the intervention with Group 1 while other groups remained in baseline. After 7 days, the intervention was introduced to Group 2, and after 31 more days, to Group 3. The researcher staggered the introduction of the intervention to control for maturation and history and the length of baseline was varied by group. For the pre-intervention natural baseline, Group 1 experienced five sessions, Group 2 experienced eight sessions, and Group 3 experienced 22 sessions. For the pre-intervention choice baseline, Group 1 experienced five sessions, Group 2 experienced six sessions, and Group 3 experienced seven sessions. The researcher used a changing criterion design for the intervention phase and demonstrated behavior changes to specific levels of gradually changing response criterion for reinforcement.
4.1.6 Interobserver Agreement and Treatment Fidelity

A second observer independently collected data for the purpose of assessing interobserver agreement (IOA) in each natural baseline, choice baseline, and intervention phase. The researcher calculated the total count IOA. For natural baseline, the researcher divided the smaller count divided by the larger count multiplied by 100. For choice baseline and intervention, the researcher calculated trial-by-trial IOA. The researcher divided the number of LLR selection in agreement by the total number of LLR selection in agreements and disagreements and multiplied by 100. The researcher obtained IOA data for 50% of natural baseline sessions, 38% of choice baseline sessions, and 26% of intervention sessions with 100% agreement. A second scorer also independently collected data for the purpose of assessing IOA for the BRIC Task using a trial-by-trial IOA. The researcher divided the number of trials in agreement by the total number of trials in agreements and disagreements and multiplied by 100. The researcher obtained IOA data for 53% of sessions with 100% agreement.

The researcher collected treatment fidelity data in each natural baseline, choice baseline, and intervention phase. For the natural baseline phase, a second independent observer recorded whether the researcher (a) posted the worksheet at 9 a.m. and (b) removed the worksheet at 2 p.m. For the choice baseline and intervention phase, a second independent observer recorded whether the researcher (a) the researcher provided the correct number of choices, (b) provided the correct number of math facts, and (c) provided the correct consequence contingent on the choice. The researcher obtained fidelity data for 35% of the natural baseline sessions, 23% of the choice baseline sessions, and 21% of the intervention sessions. Fidelity was 100% across all sessions and conditions.

4.2 Results
4.2.1 Individual Participants

Figure 1 displays the percentage of LLR selection for individual participants in Group 1 (top panel), Group 2 (middle panel), and Group 3 (bottom panel). Grey bars represent pre-intervention choice responding and black bars represent post-intervention choice responding. Out of the seven participants in Group 1 (on grade-level), six participants showed low level of LLR selection prior to the intervention (top panel). After the intervention, four participants showed increased level of LLR selection, one participant showed decreased level of LLR selection, and two participants remained constant at 100% LLR selection. Out of the six participants in Group 2 (below grade-level), all six participants showed low level of LLR selection prior to the intervention (middle panel) during the choice baseline. After the intervention, three participants showed increased level of LLR selection, one participant showed decreased level of LLR selection, and two participants remained constant. Out of the initial seven participants in Group 3 (above grade-level), two participants, Participant H and N, moved classrooms during the intervention and were removed from the final analysis. Thus, the researcher analyzed the remaining five participants (bottom panel). Out of the five participants, two participants emitted 100% LLR selection during the choice baseline prior to the intervention and three participants showed low level of LLR selection. After the intervention, one participant showed increased level of LLR selection, three participants showed decreased level of LLR selection, and one participant remained constant at 100% LLR selection. Collectively, the majority of participants in groups 1 and 2 increased their selection of LLR following intervention (7 of 13, 54%) or responded similarly to baseline (4 of 13, 31%). However, this effect was not observed with group 3 (above grade-level) individual participants. It is important to note that this may be because students performing above grade-level in math generally has higher level of self-regulation (St.
Clair-Thompson & Gathercole, 2006), suggesting that the students in Group 3 may not have needed the intervention. Thus, consistent with past research, the effects would not have been observed with students in Group 3.

Figure 2 shows the number of math problems solved in the last three days of pre-intervention natural baseline, intervention, and post-intervention natural baseline per participant. The last three days were analyzed to examine data after sufficient exposure to a condition (baseline, intervention, and post-intervention). During the pre-intervention natural baseline phase, 13 participants completed 0 math problem, four participants completed an average of 11 math problems, and one participant completed 100 math problems across all three days. Once the intervention was implemented, 17 students showed an increased level of number of math problems solved compared to the pre-intervention natural baseline phase. One participant showed a decreased level of number of math problems solved during the intervention compared to the pre-intervention natural baseline phase. When the researchers withdrew the intervention and returned to natural baseline condition, 16 participants showed a decreased level of math problems solved and two participants showed an increased level. Collectively, the participants completed more math problems during the intervention phase compared to conditions without the intervention since the majority of the participants (15 of 18, 83%) showed higher number of math problems solved during the intervention phase than the pre- and post-intervention natural baseline phase.

4.2.2 Group of Participants

Figure 3 displays the percentage of LLR selection for Group 1 (top panel), 2 (middle panel), and 3 (bottom panel), respectively. These represent each group’s percentage of LLR selection for each session. The researcher obtained the number by dividing the number of
students selecting LLR by the number of all students in the respective group per session, multiplied by 100. Before the intervention, an average of 38% (range, 29 to 43) of the participants in Group 1 (on grade-level) selected LLR for each trial across five sessions. After the intervention, the average increased to 70% (range, 33 to 86) across five sessions. For Group 2 (below grade-level), an average of 36% (range, 0 to 60) of the participants selected LLR across nine sessions prior to the intervention. After the intervention, the average increased to 57% (range, 40 to 75) across nine sessions. For Group 3 (above grade-level), prior to the intervention, an average of 65% (range, 50 to 71) of participants selected LLR selection across six sessions. After the intervention, the mean decreased to 50% (range, 25 to 75) across five days. This also may be because students in Group 3 did not need the intervention.

Figure 4 shows the mean number of math facts completed by Group 1 (top panel), Group 2 (middle panel), and Group 3 (bottom panel). Each data point represents the average number of math facts completed for a group and the data points represent the condition (e.g., when a group chose the SSR during choice baseline, the data point is an “x” to represent this condition and the data point is at zero math facts because this condition requires no completion of math facts). All three groups showed a decreasing trend of the mean number of math facts solved in the natural baseline phase after an increase in the first 2 days. During the choice baseline prior to the intervention, Group 1 (on grade-level) selected the LLR across 0 out of 5 days. Once the group met criterion after 26 days of the intervention by completing 50 math problems to gain access to 10 min of preferred activities, the group returned to choice baseline and selected the LLR across 4 out of 5 days. The natural baseline shows a moderate level with high variability (top panel). Before the intervention during the choice baseline, Group 2 (below grade-level) selected the LLR across 3 out of 9 days. The group met criterion after 30 days of the intervention and returned to
choice baseline phase. After the intervention, Group 2 (below grade-level) selected the LLR across 5 out of 9 days and the natural baseline phase showed decreasing trend similar to the pre-intervention natural baseline phase (middle panel). During the choice baseline before intervention, Group 3 (above grade-level) selected the LLR across 5 out of 6 days. Once the group met criterion across 26 days, the group selected the LLR across 3 out of 5 days and the natural baseline phase showed decreasing trend similar to the pre-intervention natural baseline phase (bottom panel).

4.2.2.1 Statistical Analysis

Analyses of 18 students focused on the effects of the intervention on the participants’ LLR selection and BRIC Task score. The researchers conducted the BRIC Task assessment during the pre- and post-intervention natural baseline phases. Overall, the results of Group 1 and 2 (on and below grade-level) showed increased LLR selection ($M_{\text{diff}} = 27.462$, $SD = 45.990$) from baseline to intervention, while Group 3 (above grade-level) showed decreased level of LLR selection ($M_{\text{diff}} = -4.600$, $SD = 42.081$). Similarly, Group 1 and 2 (on and below grade-level) showed improved indifference point on the BRIC Task after the intervention ($M_{\text{diff}} = 4.077$, $SD = 40.289$) while Group 3 (above grade-level) showed decreased indifferent point on the BRIC Task ($M_{\text{diff}} = -46.200$, $SD = 43.304$). Given the different directionality of the results, Group 1 and 2 and Group 3 were analyzed separately.

A dependent samples t-test was conducted to determine if the difference between the participants’ pre- and post-intervention level of LLR selection and indifferent point on the BRIC Task were significantly different. For Group 1 and 2 (on and below grade-level), the baseline and post-intervention levels of LLR selection were significantly different, $t (12) = 2.153, p = .052; d = 0.597$. There was medium effect size ($d = 0.597$), suggesting that despite the no significance,
the training was moderately effective at increasing LLR selection for Group 1 and 2. However, the pre- and post-intervention BRIC Task indifference point were not significantly different and there was a small effect size, $t(12) = .365, p = .722; d = .101$. For Group 3 (above grade-level), there was no significant difference between pre- and post-intervention levels of LLR selection with a small effect size, $t(4) = -.244, p = .819; d = -.109$, or the pre- and post-intervention BRIC Task indifference point were not significantly different but had a large negative effect size, $t(4) = -2.386, p = .076; d = -1.067$. In summary, although there was no statistical difference, the intervention had a moderate effect on increasing LLR selection for Group 1 and 2 (on and below grade-level) but not for Group 3 (above grade-level). Changes in the BRIC Task scores were not significant across all three groups but the direction of changes remained similar to the results of LLR selection – the score increased after the intervention for Group 1 and 2 (on and below grade-level) while it largely decreased for Group 3 (above grade-level).

4.3 Discussion

The purpose of this study was to evaluate the effects of a progressive delay training on students’ advantageous choice-making behavior. Specifically, we measured the students’ selection of LLR over SSR and BRIC Task score before and after the intervention to determine the students’ advantageous choice-making behavior in an educational setting. While previous evaluations of this intervention targeted individuals on a one-to-one basis, we extended this to a group context during online schooling. We found that, even though statistically not significant, the intervention had a moderate effect size at increasing LLR selection for Group 1 and Group 2 (on and below grade-level in math). This means that these students demonstrated greater selection of LLR after the intervention. On the other hand, Group 3 (highest math level) showed a decrease in LLR selection even though the decrease was not statistically significant. The
indifference point derived from the BRIC Task, even though statistically not significant, increased for Group 1 and 2 while decreased largely for Group 3.

The findings have implications for practice and literature on advantageous choice-making behavior of younger children. First, this study extends the literature showing the effectiveness of progressive delay of self-control training (Dixon & Falcomata, 2004; Schweitzer & Sulzer-Azaroff, 1988) and progressive delay training with choice opportunities (Dixon & Tibbetts, 2009), by implementing the intervention on a group basis during online learning. The majority of past studies successfully changed the target behavior when the intervention was implemented for individual participants. The current study implemented the intervention on a group level to determine the feasibility of the intervention with multiple participants in a general educational setting. Thus, this study directly addressed the future suggestions offered by Staubitz et al. (2020) by implementing the progressive delay training procedure to a general education population. The findings in the current study suggested that the intervention can be delivered to multiple individuals at the same time and change the target behavior effectively, highlighting the cost-effectiveness and efficiency of the intervention. This is especially important for educators who have limited resources and time in an educational setting and want to maximize the efficiency of the instruction delivered.

One thing to note is that the intervention was not effective for students performing above-grade level for math. This may be because students performing above grade-level for math are more likely to demonstrate strong self-regulation skills (St. Clair-Thompson & Gathercole, 2006) and thus, the students in Group 3 may not have needed the intervention. The BRIC Task scores showed that the mean pre-intervention subjective value of $100 at 1 month was highest for Group 3 (66.140; above grade-level) compared to Group 1 (66.000; (on grade-level) and Group 2
(39.57; below grade-level)), indicating that above grade-level students in Group 3 discounted the least out of the three groups, suggesting that they engaged in more advantageous choice-making behavior than the students in the other groups. Several studies have shown that having strong self-regulation skills is related to school success in both math and reading (Adams & Snowling, 2001; Connor et al., 2010; Day et al., 2015; St. Clair-Thompson & Gathercole, 2006). Since the above grade-level students already demonstrated a higher score on the BRIC Task (i.e., engaging in more advantageous choice-making behavior), there might have been a ceiling effect at increasing the target behavior. Another explanation is that the intervention may have functioned as punishment for those who already demonstrated greater advantageous choice-making behavior. The participants in Group 3 (above grade-level) consistently chose the LLR in the pre-intervention choice baseline phase but experienced the intervention which started from providing reinforcement for completing in no extra math facts. In turn, the effectiveness of the reinforcer may have been decreased and the frequency of choosing the LLR decreased. In the future, educators should evaluate whether individuals require an intervention to promote advantageous choice-making behavior prior to implementation and only target individuals requiring an intervention and/or performing at on or below grade-level on academics.

Second, the results inform educators on how to teach advantageous choice-making behavior for younger population, specifically those who perform below or on grade-level for math. In line with past research (Fisher et al., 1997; Mazur & Logue, 1978; Schweitzer & Sulzer-Azaroff, 1988), gradually increasing delays to the larger reinforcement with choice opportunities improved delay of gratification. Educators should incorporate progressive delay and choice-making opportunities to instruction or token reinforcement system to provide increased opportunities to teach advantageous choice-making behavior throughout the day. Further, the
study lends support to the existing literature on the use of concurrent activity during the delay to reinforcement since the participants engaged in solving math facts during the delay. Studies showed that having individuals engage in concurrent activity during the delay rather than having them simply wait for the reinforcer can improve advantageous choice-making behavior (Dixon & Cummings, 2001; Dixon & Holcomb, 2000). This means that educators could provide an activity for students during a waiting period to promote advantageous choice-making behavior.

Educational benefits also follow the implementation of the intervention procedure to instruction. Most participants in this study increased the number of practice math problems completed during the intervention phase compared to pre- and post-natural intervention phases. The finding suggested that the participants received more opportunities to practice math fact fluency, which is in line with the common core standard (2.OA.B.2) and is the foundational skill for other standards such as fluently adding and subtracting within 100 (2.NBT.B.5) and within 1,000 (2.NBT.B.7). Educators are encouraged to incorporate a similar procedure during math and potentially other subject areas to efficiently provide greater learning opportunities.

Third, the findings suggest that verbal behavior may relate to advantageous choice-making behavior. Skinner (1957) argued that delayed consequences are often controlled by verbal statements of a contingency or rules, which are contingency-specifying stimuli produced by one’s own verbal behavior (Zettle, 1990). To have delayed consequences reinforce behavior, an individual must be able to hear oneself say the rules with specific contingencies. When an individual functions as a speaker and a listener (i.e., speaker-as-own-listener; Greer & Ross, 2008), this individual can recite the rules to him or herself and have greater tolerance for delayed consequences. Given the possible implications, future studies should investigate the relation
between verbal behavior and advantageous choice-making behavior using delay discounting measures.

The study is not without limitations. One limitation was experimental control. Typically for greater experimental control in a multiple baseline across groups design, the participants in each group should be matched by their chronological age, academic level, or any relevant variable that may affect participants’ responding to the intervention. However, the current study grouped the participants by their academic levels in math rather than matching. The participants were grouped based on their math levels to better target participants requiring the intervention the most. Past studies showed that academic success in math related to self-regulation (St. Clair-Thompson & Gathercole, 2006), indicating that students performing below grade-level in math were more likely to have lower level of advantageous choice-making behavior and would need the intervention the most. Consistently, the baseline data showed that the students in Group 2 (below grade-level) showed the lowest percentage of LLR selection (36%) while Group 3 (above grade-level) showed the highest percentage of LLR selection (65%). As a result, the intervention was effective at increasing LLR selection for students performing on and below grade-level while it was not for those performing above grade-level in math. While the study did not match the groups, the study still showed experimental control using staggered introduction of the intervention, repeated measures from baseline, verification of the intervention changing Group 1 and 2’s responding, and replication across Group 1 and 2 (Horner et al., 2005). To better understand the general effects of the intervention across participants with different academic levels, future researchers should examine the effects of the intervention when the groups are matched based on the students’ level of academic skills.
A second limitation is the small sample size. Group 1 and 2 (on and below grade-level) only consisted of 13 students and Group 3 (above grade-level) only consisted of seven students of which two students were excluded because of midyear classroom changes. Due to the small sample size, there was low statistical power which reduces the chance of detecting a potential effect. The current study addressed this by reporting and interpreting the effect sizes to explain the effects of the intervention. Nonetheless, for a more accurate statistical effect of the intervention, future studies should have a larger number of participants.

Third, the outcomes of this classwide study may have been peer influenced by the group nature of the intervention. Based on the beginning of the year assessment, all the participants had observational learning in repertoire, which means that the participants were able to observe and learn from each other’s behavior (Greer et al., 2006). Although the researcher ensured that the students closed their eyes to vote, the participants could hear each other talk about their choices before and after the choice, which may have affected the students’ choices during baseline and post-intervention with unknown effects on the data. Future studies should control peer influence on choice-making behavior or directly measure the varying degree of peer influence on the participants to control for analysis. Researchers should also examine how different modalities (e.g., anonymous voting) affect choices.

Last, the current study only measured the advantageous choice-making behavior in the context of math. However, studies showed association between strong self-regulation and school success in reading (Adams & Snowling, 2001; Day et al., 2015), which means that progressive delay training in conjunction with choice opportunities can be used to gauge and enhance advantageous choice-making behavior across academic subjects including reading and math. The intervention can be further extended to social skills and self-management. For example, future
studies can teach children to initiate and maintain cooperative play with a peer by using this procedure. The researcher can progressively increase the time the participant engages in conversation with a peer while providing choices of playing with a friend and earning the LLR and playing by oneself and earning the SSR. These types of studies would give light to generalizability of the intervention across different applied domains.

Despite the limitations, the present study showed that the progressive delay training was effective at increasing advantageous choice-making behavior for individuals performing on and below grade-level for math. The intervention can be an efficient and cost-effective way to teach younger students to engage in advantageous choice-making behavior in an educational setting. The finding may serve as a foundation for understanding choice-making behavior in younger population and in turn, for the development of interventions that can effectively alter this socially significant behavior.
References


https://doi.org/10.1080/17470210500162854


https://doi.org/10.1901/jaba.2006.158-04


https://nces.ed.gov/surveys/sass/tables/sass0708_035_s1s.asp


https://doi.org/10.1177/0022219419851750


https://doi.org/10.1111/j.1469-7610.2010.02347.x
Table 1

Participant Demographic Information

<table>
<thead>
<tr>
<th>Group 1 (On Grade-Level)</th>
<th>Participant</th>
<th>Age</th>
<th>Gender</th>
<th>Race/Ethnicity</th>
<th>IEP</th>
<th>ELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>7.25</td>
<td>F</td>
<td>W</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>P</td>
<td>7.75</td>
<td>M</td>
<td>W</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>MT</td>
<td>7.25</td>
<td>M</td>
<td>W</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>VK</td>
<td>8.25</td>
<td>F</td>
<td>W</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>V</td>
<td>7.42</td>
<td>F</td>
<td>MR</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>K</td>
<td>7.92</td>
<td>M</td>
<td>A</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>D</td>
<td>7.67</td>
<td>M</td>
<td>H</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 2 (Below Grade-Level)</th>
<th>Participant</th>
<th>Age</th>
<th>Gender</th>
<th>Race/Ethnicity</th>
<th>IEP</th>
<th>ELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>7.92</td>
<td>F</td>
<td>H</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>A</td>
<td>7.67</td>
<td>F</td>
<td>B</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>VL</td>
<td>7.50</td>
<td>F</td>
<td>B</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>J</td>
<td>7.42</td>
<td>M</td>
<td>B</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>T</td>
<td>7.17</td>
<td>M</td>
<td>B</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>JN</td>
<td>7.33</td>
<td>M</td>
<td>H</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 3 (Above Grade-Level)</th>
<th>Participant</th>
<th>Age</th>
<th>Gender</th>
<th>Race/Ethnicity</th>
<th>IEP</th>
<th>ELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>7.58</td>
<td>M</td>
<td>W</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>M</td>
<td>7.75</td>
<td>F</td>
<td>W</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>S</td>
<td>7.42</td>
<td>F</td>
<td>MR</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>C</td>
<td>7.58</td>
<td>M</td>
<td>H</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>G</td>
<td>7.83</td>
<td>F</td>
<td>H</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>H</td>
<td>7.92</td>
<td>F</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>8.08</td>
<td>M</td>
<td>W</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

**Note.** IEP = Individualized Education Program; ELL = English Language Learner; F = Female; M = Male; W = White; MR = Multi-Racial; A = Asian; H = Hispanic; B = Black; Y = Yes; N = No.
Figure 1

Percentage of Larger Later Reinforcer (LLR) Selection per Participant

Note. The number on top of each bar represents the number of LLR selection over the total days of selection opportunities.
Figure 2

Number of Math Problems Solved in the Last Three Days of Pre-Intervention Natural Baseline, Intervention, and Post-Intervention Natural Baseline per Participant.
Figure 3

Percentage of Larger Later Reinforcer (LLR) Selection

Note. The dotted line represents the mean percentage of LLR selection of each phase.
Figure 4

Mean Number of Math Facts Completed for Group 1, 2, and 3

Note. NBL = Natural Baseline; SSR = Smaller Sooner Reinforcer; LLR = Larger Later Reinforcer; P-LLR = Progressive Larger Later Reinforcer.
Appendix A

An Example of Slides Displayed for Choice Baseline Sessions

No more math!
2 minutes of mystery prize

50 math facts
10 minutes of mystery prize
Appendix B

An Example of Slides Displayed during Delay of Gratification Training Phase

No more math!
2 minutes of mystery prize

5 math facts
10 minutes of mystery prize

50 math facts
10 minutes of mystery prize
Chapter 5: Delayed Consequences in General Education through the Lenses of Verbal Behavior and Delay Discounting

Outline

Delayed consequences have important academic, behavioral, and health implications for younger children. While behavior analysts have investigated rule-governed behavior within the context of delayed reinforcement, no studies have analyzed the potential correlations among variables contributing to tolerance for delayed consequences. Studies suggested that concepts related to delayed consequences within behavior analysis are verbal behavior and delay discounting. Given the educational importance of delayed consequences, the researchers studied the relationship among verbal behavior repertoires, delay discounting, and academic achievement in a second-grade general education classroom. Specifically, the researchers used a hypothetical binary choice task to derive delay discounting measures and assessed the accurate responses to the Incidental Bidirectional Naming (Inc-BiN) probes as a verbal behavior measure. The results showed that there was a significant positive relation between delay discounting measures and accuracy of Inc-BiN responses. There was no significant relation between academic achievement and either delay discounting or Inc-BiN. Future research on bridging the concepts of verbal behavior and delay discounting is discussed.

Keywords: delay discounting, incidental bidirectional naming, general education, verbal behavior, academic achievement.
Delayed Consequences in General Education through the Lenses of Verbal Behavior and Delay Discounting

Many human behaviors do not have an obvious immediate consequence and there are times when we act for consequences that cannot be produced immediately. Students work hard to receive a final grade often received at the end of a semester, adults work to receive a biweekly paycheck, and people work out regularly to improve health. Despite the clear definition of a consequence, which is change in the environment following a behavior that makes the behavior occur more or less frequently in the future (Cooper et al., 2019; Skinner, 1953), there has been less consensus on the concept of delayed consequences. Delayed consequence is a general term for when there is a delay between the response and the consequences. Malott (2008) defined “immediate” as no longer than 1-s delay and defined “delayed” as between 1 s and 60 s duration between the behavior and consequence. He argued that if the reinforcer is delivered after 60 s, it is likely that the response will not be reinforced. Nevertheless, several studies showed that organisms acquire and maintain responding with delayed consequences (Chritchfield & Lattal, 1993; Sutphin et al., 1998; Wilkenfield et al., 1992; Williams & Lattal, 1999). A delayed consequence is indeed a consequence because it is part of a causal chain and organisms can acquire and maintain responding contingent on receiving a delayed reinforcer.

Tolerance for delayed consequences is important for school-aged children. Past studies showed that children who can wait for a more advantageous reinforcer available later over a smaller, immediately available reinforcer performed better academically, were more verbally fluent, attentive, planful, and able to handle stress better as adolescents (Mischel et al., 1988). Children who have difficulty waiting for the larger, more advantageous reinforcer made less academic progress and experienced more anxiety, depression, and aggressive behavior problems.
(McClelland et al., 2007; Martel et al., 2007). Long term implications included higher risk for poor health outcomes including obesity and drug dependency (Moffitt et al., 2011). Academically, numerous studies showed that students who wait longer for a larger reinforcer have greater success in both math and reading (Adams & Snowling, 2001; Bembenutty, 2011; Connor et al., 2010; Day et al., 2015; St. Clair-Thompson & Gathercole, 2006). Collectively, the studies highlight the academic, behavioral, and health implications of delayed consequences for younger children.

A concept within behavior science that studies the effects of delayed consequences is delay discounting, which is a concept within behavioral economics that examines preferences between smaller, sooner and larger, later reinforcers (SSR and LLR, respectively). The general finding in this area is that a reinforcer’s value diminishes across time. However, researchers have studied individual differences in delay discounting and revealed relations such as children discount quicker than adults (Steinberg et al., 2009) and smokers discount quicker than non-smokers (Mitchell, 1999). Delay discounting has been studied extensively with both human and nonhuman subjects (e.g., Vanderveldt et al., 2016), using a variety of procedures, ranging from hypothetical choice arrangements to experiential tasks (Reed et al., 2020). All discounting tasks are characterized by the measurement of organisms’ responding towards either SSRs or LLRs. For example, a series of question might entail a participant choosing between a fixed amount of money today verses a larger amount of money 1 day, 1 week, 1 month, or 1 year later. From this assessment, researchers map indifference points, which are points at which the value switches from the LLR to the SSR (i.e., the points at which the subjective values of LLR and SSR are equal; Reed et al., 2013), to quantify tolerance to delayed consequences. Studies have shown that delay discounting is related to academic achievement in younger populations. In a study with
adolescents, students with lower levels of temporal discounting performed better academically (Lee et al., 2012). Further, there was significant association between delay discounting and educational attainment of Irish teenagers (i.e., grades on exams; Freeney & O’Connell, 2010). Clear academic implications of delay discounting further emphasize the importance of younger population’s tolerance for delayed consequences.

Another concept within behavior science that can help explain tolerance for delayed consequences is verbal behavior. Verbal behavior plays a significant role in explaining delayed consequences because they are often under the control of rules. Skinner (1957) argued that delayed consequences are often controlled by verbal statements of a contingency (i.e., rules). Studies on rules and delays showed that the delay itself minimally controlled behavior. Rather, the deadline and timing of reinforcement (i.e., specific contingencies), controlled behavior. In other words, specification of rules including due dates/times and whether the reinforcement will be delivered immediately or delayed increased responses. Braam and Malott (1990) investigated this directly by comparing the control exerted by different types of rules on the task completion rate of preschoolers. Specifically, the researchers compared six different conditions consisting of combinations of the presence or absence of a deadline and immediate, delayed, or no reinforcer across two different types of tasks. They showed that rules specifying immediate deadline with immediate delivery of the reinforcer and immediate deadline with one-week delay in the delivery of the reinforcer exerted reliable control over task completion (i.e., highest level of task completion). On the other hand, specifying only response requirements (i.e., no deadline nor contingency for reinforcer) or only the delay with no deadline did not reliably control behavior (Braam & Malott, 1990).
Rules are contingency-specifying stimuli produced by one’s own verbal behavior (Zettle, 1990). To have delayed events function as reinforcement, one must be able to hear oneself recite the rules with specific contingencies. This means that one must be able to recite the rule as a speaker and listen to the rules within one’s own skin. Then, the rules and specific contingencies would signal the availability of reinforcement but would not directly reinforce the behavior. Reinforcement is rather created through the correspondence between rehearsing the rules (“saying”), listening to oneself, and following them (“doing”). In verbal behavior, this correspondence between saying and doing is referred as say-do correspondence, which is defined as “a verbal behavior cusp that enables one to function as a listener to their own verbal behavior and follow the directions given to him/herself” (Greer et al., 2017, p. 686). In other words, an individual can listen to one’s own speaking of what they will do in the future and perform the behavior. For example, a person can say, “I will finish writing a paper and then go watch one episode of Seinfeld.” This person has say-do correspondence if he or she starts to write the paper and watches one episode of Seinfeld once done writing. He or she emitted an instance of verbal behavior and performed the behavior. With the onset of this repertoire, an individual functions as both a speaker and listener or as a speaker-as-own listener (Greer & Ross, 2008). This means that this individual starts to respond as a listener and provide for their own speaker behavior, which function as the precursor for thinking, self-awareness, and problem solving (Greer & Ross, 2008; Greer & Keohane, 2006).

Speaker-as-own-listener cusps are necessary for an individual to become truly verbal (Greer & Keohane. 2005; Greer & Ross, 2008). Verbal Behavior Development Theory (VBDT; Greer & Keohane. 2005; Greer & Ross, 2008) extended Skinner’s verbal behavior (1957) by outlining a developmental trajectory of listener and speaker repertoires and identifying speaker-
as-own-listener cusps necessary for becoming truly verbal. Acquisition of cusps allow individuals to contact new contingencies in the environment (Rosales-Ruiz & Baer, 1997), while cusps that are also capabilities allow individuals to learn in new ways that they could not have before (Greer & Ross, 2008). Say-do correspondence, along with self-talk, are behavioral cusps that are speaker-as-own-listener repertoires while Incidental Bidirectional Naming (Inc-BiN) is identified as a speaker-as-own-listener cusp that is also a capability (Greer & Keohane, 2005; Greer & Ross, 2008; Greer et al., 2017; Horne & Lowe, 1996; Skinner, 1957).

Similar to say-do correspondence, Inc-BiN indicates joint stimulus control for speaker and listener repertoires, highlighting that Inc-BiN may be essential in explaining individuals’ tolerance for delayed consequences. Inc-BiN is a capability that combines conventional speaker and listener behavior within an individual (Horne & Lowe, 1996; Skinner, 1957). When an individual has the stimulus control for Inc-BiN, speaker and listener behaviors are under a joint stimulus control such that models of object-name relations jointly control both speaker and listener behavior (Greer, 2020). This individual can function as a speaker and listener and incidentally learn language from his or her environment, resulting in faster rate of learning (Greer & Ross; 2008; Hranchuk et al., 2019). Greer (2020) argued that if an individual acquires stimulus control for listener and speaker verbal operants (i.e., more complex Inc-BiN stimulus control), the degree of reinforcement value changes for that stimulus and thus the stimulus selects out his or her behavior. This means that the degree of joint stimulus control of speaker and listener behavior (i.e., degree of Inc-BiN) affects how rules control an individual’s behavior. When a person has higher degree of Inc-BiN, this person is more likely to have stronger joint stimulus control for speaker and listener behavior, resulting in rules having stronger stimulus control and rules more likely to controlling one’s behavior. Thus, this person would have greater
tolerance for delayed consequences. Different levels of stimulus control of rules can account how delays affect individuals differently.

Delayed consequences have implications on delay discounting, verbal behavior, and learning. However, to the best of our knowledge, nobody has investigated the correlations between delay discounting and verbal behavior. Hence, the primary goal of the paper was to determine the relationship between the accuracy of Inc-BiN responses and delay discounting. Thus, the researchers asked the following research question: (a) will students with higher accuracy of Inc-BiN responses have lower rate of discounting (i.e., discount the subjective value at a slower rate, indicating an individual more frequently selecting the LLR over SSR). As a secondary goal, the researchers investigated the relationship between the accuracy of Inc-BiN responses and delay discounting and academic achievement. Therefore, the researchers asked the following additional research questions: (b) will students with higher accuracy of Inc-BiN responses have higher academic achievement, and (c) will students with lower rate of discounting have higher level of academic achievement?

5.1 Method

5.1.1 Participants and Setting

Twenty-three first and second graders (10 female and 13 male participants) participated. The participants’ ages ranged from 6.08 to 7.92 (\(M = 7.01, SD = 0.59\) years) years old. Six participants had Individualized Education Programs (IEP) with classifications of specific learning disability and speech or language impairment. Of the 23 participants, five participants qualified as English Language Learners (ELL). The sample consisted of students who were White (61%), Hispanic (30%), Black (4%), and multiracial (4%). All participants had the
prerequisites to sit appropriately, log on to a computer, complete an assessment on the computer independently.

The study took place in a Title 1 public elementary school located in a suburb outside of a large metropolitan area with grades from pre-school to second. All participants were in a first- or second-grade classroom that used a Comprehensive Application for Behavior Analysis to Schooling (CABAS®, Accelerated Independent Learner program; Greer, 1994) model. Each first- and second-grade classroom consisted of 15 students, one head teacher, and two teaching assistants. The model incorporates a scientific approach to pedagogy, learning, curriculum, and classroom management, and all instruction is individualized to each student’s repertoire. All data were collected in the participants’ classroom throughout the day during noninstructional periods. The researchers conducted the delay discounting assessment and bidirectional naming probes at individual tables in the classroom during the school day from October to January. Beginning of the year i-Ready® K-12 Adaptive Reading Diagnostic (Curriculum Associates, 2021) were conducted at individual tables in the classroom in September of the same year.

5.1.2 Materials

For the BiN assessments, the researcher used Microsoft PowerPoint slides and a data sheet. There were three probe sets. Each set consisted of five unfamiliar stimuli (i.e., symbolic and purely intraverbally related stimuli) which were assessed with listener and speaker probes. For the listener probes, each PowerPoint slide consisted of a target stimulus in a field of three. For the speaker probes, each PowerPoint slide consisted of one of the five target unfamiliar visual stimuli. Set 1 of unfamiliar stimuli included square root of (\(\sqrt{}\)), braces (\(\{\} \)), integral (\(\int\)), pi constant (\(\pi\)), and ohm sign (\(\Omega\)). Set 2 of unfamiliar stimuli included sodium (Na), iron (Fe),
silver (Ag), mercury (Hg), and antimony (Sb). Set 3 of unfamiliar stimuli included congruent ($\cong$), sum of ($\Sigma$), empty set ($\emptyset$), delta ($\Delta$), and membership ($\in$).

For the delay discounting assessment, the researcher used Microsoft PowerPoint slides and a data sheet (Appendix A; from Critchfield & Atteberry, 2003 and Reed & Martens, 2011). Each PowerPoint slide consisted of a written direction, “Would you rather have,” on the top of the screen and had two boxes (7.62 cm X 7.62 cm) underneath it with the two reward values in each box. Each box represented a SSR value and a LLR monetary value.

The school district provided the i-Ready assessments, which is an online screening tool students take three times a year to capture a complete picture of student performance and growth (Curriculum Associates, 2021). The diagnostic assesses a wide range of skills including those above and below a student’s chronological grade and the assessment questions adapt to student responses. Based on the results, personalized instruction is generated to help educators address strengths and needs of students. i-Ready reading and mathematics diagnostics assessment obtained the highest rating for reliability and validity according to the Academic Progress Monitoring Tools Chart by National Center on Intensive Intervention (2021). All students used a Google Chromebook and headphones provided by the district and completed the assessment using their district login information.

5.1.3 Measurement

5.1.3.1 Incidental Bidirectional Naming (Inc-BiN)

5.1.3.1.1 General Procedure

The researcher conducted the probe sessions in the order of Set 1, Set 2, and Set 3. For each set, the researcher started with the naming experience and then conducted the probes. The researcher repeated the procedure for Set 2 and then Set 3.
5.1.3.1.2 Naming Experience

The researcher presented a novel set of five unfamiliar visual stimuli for 20 total presentations. Each slide contained one target visual stimulus. Once the researcher gained the participant’s attention, they provided the vocal antecedent, “I am going to show you the names of symbols.” The researcher provided the name for each stimulus through a vocal tact. The participant had a total of 20 opportunities to observe the target stimuli for each unfamiliar stimuli set.

5.1.3.1.3 Inc-BiN Probes

Researchers measured the number of correct responses to untaught Inc-BiN visual stimuli two hours following the Naming experience. These probe sessions were unconsequated. There were 10 trials that consisted of listener responses and 10 trials of speaker responses. For the listener responses, researchers presented the target stimulus along with two negative exemplars. Researchers provided a vocal antecedent of, “Which one is ___?”. The participant pointed to a stimulus following the vocal antecedent. For the speaker responses, there was a total of 10 intraverbal tacts. The researcher provided the vocal antecedent, “What is this?”. Researchers rotated stimuli presentations across the five target stimuli, so no same stimuli were presented consecutively. A correct response was defined as pointing to the corresponding stimulus or tacting the name of the corresponding stimulus within 3 s. An incorrect response was defined as pointing to an incorrect stimulus or tacting an incorrect name or no response. Inc-BiN was present in a participant’s repertoire if the participant emitted at least 80% accuracy in both listener and speaker responses for two of the three sets of novel stimuli. For this study, the researcher added the number of correct listener and speaker responses across the three sets. There was a total of 30 listener and 30 speaker responses across the three sets.
5.1.3.2 Delay Discounting Assessment

The researcher administered the delay discounting assessment to obtain the indifference points across eight delays (i.e., 1 day, 5 days, 1 month, 2 months, 6 months, 9 months, 1.5 years, and 4 years). For each delay, there were 22 possible indifference points (Appendix A; from Critchfield & Atteberry, 2003 and Reed & Martens, 2011). The researcher used hypothetical, monetary choices of SSR and LLR and asked each participant a series of questions with binary choices. For each question, the participant was required to choose one of two monetary choices that had a contrasting length of delay and size of reinforcer. The researcher manipulated the amount of SSR values while holding the LLR delay value constant to determine the points at which each participant changed his or her choice from the LLR to the SSR.

All participants started with the shortest delay (i.e., 1 day) and increased in the order of delay until they reached the longest delay (i.e., 4 years). For each delay, the researcher first presented a binary choice between SSR of $50 and LLR of $100. The researcher provided the vocal antecedent, “Would you rather have $50 right now or $100 in one month?” If the participant chose the $50 available now, the next trial consisted of a smaller amount of money available now against $100 available in one month. However, if the participant chose $100 in one month, the next trial consisted of a larger immediate monetary amount and the delayed $100. The researcher collected data on a data sheet with all possible choices (Appendix A; from Critchfield & Atteberry, 2003 and Reed & Martens, 2011). Then, the researcher derived the subjective value of the $100 (i.e., indifference point) in terms of smaller amount of money available immediately (see Critchfield & Atteberry, 2003; Reed & Martens, 2011), the $k$ value (i.e., free parameter that represents the degree of discounting based on the steepness of the curve; Odum, 2011) and area under the discounting curve (AUC; Myerson et al., 2001). Subjective
value of the $100 was used because it has been used to predict students’ classroom behavior to measure sensitivity to reward delays (Kim et al., 2022). Higher subjective value of $100 at 1-month delay indicated greater tolerance for delayed consequences. Higher $k$ value suggested higher degrees of discounting or less tolerance for delayed consequences. Adding the areas of a series of trapezoids under the discounting curve yields the AUC value, and higher AUC suggested greater tolerance for delayed consequences.

5.1.3.3 i-Ready Diagnostic Assessment

The current study used the national percentile that was derived from the score. The percentiles represent a nationally representative students in the same grade-level who took the diagnostic assessment at the same time of the year. For example, if a student’s percentile is 90 for a fall diagnostic assessment, he or she scored better than 90% of a nationally representative group of students who took the fall diagnostic. As of 2020, i-Ready program was used across over eight million users nationwide, which was approximately 25% of all K-8 students across all 50 states. As of 2021, i-Ready was being used across more than 14,000 schools with over 10 million students (Curriculum Associates, 2021). i-Ready is a web-based adaptive diagnostic where the difficulty of the assessment questions matches the ability of each student. The diagnostic begins with questions that are at a difficulty level that matches their chronological grade-level. When the students answer questions correctly, the test gets more difficult. When the students answer questions incorrectly, the test gets harder.

In the beginning of the year in September, all participants completed the reading and mathematics i-Ready diagnostic assessment. All students completed the assessment at their individual desks with headphones on. The students completed the math portion across two days with 30 min allotted each day and completed the reading portion across two days with 30 min
allotted each day. If students needed more time to complete the assessment, the teacher provided more time after the four days allotted for both subjects. Once the students completed the assessment, each student received a score for math and reading.

5.1.4 Statistical Analysis

The current study had a small sample size and skewed distribution. According to a Shapiro-Wilk test, the distributions significantly deviate from a normal distribution for subjective value of $100 at 1-month delay, \( W(23) = .87, p = .006 \), AUC, \( W(23) = .86, p = .004 \), and \( k \) value, \( W(23) = .53, p < .001 \). Hence, the researcher conducted a nonparametric Spearman’s rho correlations (\( r_s \)) to determine the relationship among BiN, delay discounting, and i-Ready scores.

5.1.5 Interobserver Agreement and Treatment Fidelity

A second observer independently collected data for the purpose of assessing interobserver agreement (IOA) of the delay discounting and Inc-BiN assessment. For the delay discounting assessment, the researcher calculated trial-by-trial IOA by adding the number of binary choices in agreements, dividing by the total number of binary choices of agreements and disagreements and multiplying by 100. The researcher obtained IOA data for five out of 15 BRIC tasks (33.3%) with 100% agreement.

For the Inc-BiN probes, the researcher calculated trial-by-trial IOA by adding the number of agreements, dividing by the total number of agreements and disagreements and multiplying by 100%. The researcher obtained IOA data for 33% of the probe sessions across all probe sessions with 99.7% agreement (range, 95-100%). A trained independent observer completed a Teacher Performance of Rate and Accuracy (TPRA; Ingham & Greer, 1992) form to measure treatment fidelity. The observer used a TPRA to record the accuracy of the researcher delivering (1) antecedents and (2) consequences contingent on the participant’s response. Treatment fidelity
was calculated by diving the total number of correct researcher responses by the total number of
responses and multiplying by 100. The researcher collected treatment fidelity data for 33.3% of
the probe sessions and fidelity was 100%.

5.2 Results

Table 1 shows the descriptive data and Spearman’s rho correlation among the i-Ready
percentile (reading and math), accurate responses to Inc-BiN probes, and delay discounting
measures including subjective value of $100 at 1-month delay, AUC, and $k$ value. For the
analysis, a criterion from Reed and Martens (2011) was employed to include participants
showing a discounting effect. A participant was showing a discounting effect if the mean
indifference points from the three shortest delays exceeded the mean indifference points from the
three longest delays with no more than two instances of a preference reversal (i.e., increase in
indifference point across consecutive delays). Traditionally, studies with adults allowed one
instances of preference reversal (Dixon et al., 2006). However, given the exploratory nature of
the study and younger participant age in the current study compared to previous studies (e.g.,
Reed & Martens, 2011 used sixth graders), the current study allowed two preference reversals.
Thus, participants were excluded from the analysis. In total, there were 13 out of 23 (57%)
participants included in the analysis.

There was a significantly strong relation between the degree of Inc-BiN and delay
discounting measures. Significant positive relations were found between subjective value of
$100 at 1-month delay and Inc-BiN listener responses ($r_s = .68, p = .01$) and speaker responses
($r_s = .69, p = .01$), and significant negative relations were found between the $k$ value and Inc-BiN
listener responses ($r_s = -.80, p = .001$) and speaker responses ($r_s = -.60, p = .03$). When the
subjective value of $100 at 1-month delay was higher and $k$ value was smaller (i.e., both

159
indicating greater tolerance to delayed consequences), participants showed higher level of Inc-BiN. Interestingly, there was no significant relation between AUC and Inc-BiN listener responses ($r_s = .47, p = .15$) and speaker responses ($r_s = .47, p = .11$). Thus, when using the subjective value of $100$ at $1$-month delay and $k$ value, the findings supported the primary research question that students with higher degree of Inc-BiN had lower rate of discounting.

On the other hand, the relation between delay discounting measures and i-Ready percentile was non-significant. No significant relations were found between subjective value of $100$ at $1$-month delay and i-Ready math percentile ($r_s = .32, p = .28$) and reading percentile ($r_s = .20, p = .51$), AUC and i-Ready math percentile ($r_s = -.5, p = .87$) and reading percentile ($r_s = .24, p = .44$), and $k$ value and i-Ready math percentile ($r_s = -.45, p = .12$) and reading percentile ($r_s = -.18, p = .55$). These findings did not support the secondary research question that students with lower rate of discounting had higher level of academic achievement.

Similarly, no significant relations were found between the degree of Inc-BiN and i-Ready percentiles. There was no significant relation between Inc-BiN listener responses and i-Ready math percentile ($r_s = .26, p = .39$) and reading percentile ($r_s = .00, p = .99$). Also, there was no significant relation between Inc-BiN speaker responses and i-Ready math percentile ($r_s = .31, p = .31$) and reading percentile ($r_s = .43, p = .14$). These findings did not support the secondary research question that students with higher degree of Inc-BiN had higher academic achievement.

5.3 Discussion

Tolerance for delayed consequences is important for school-aged children’s academics (Adams & Snowling, 2001; Bembenutty, 2011; Connor et al., 2010; Day et al., 2015; Mischel et al., 1988; St. Clair-Thompson & Gathercole, 2006), behavior (McClelland et al., 2007; Martel et al., 2007), and long-term health outcomes (Moffitt et al., 2011). Despite past studies suggesting
that delay discounting, verbal behavior repertoires, and learning have implications on tolerance for delayed consequences, no studies have directly investigated the relation among these variables. Therefore, the goal of this paper was to determine the association among the accurate responses to Inc-BiN probes, delay discounting, and academic achievement in younger children. The results showed that individuals who learn incidentally from their environment have greater tolerance for delayed consequences. However, academic achievement was not associated with the degree of Inc-BiN nor delay discounting.

The findings have several implications for future research on bridging the concepts of verbal behavior and delay discounting. First, the findings supported that when an individual had a higher level of Inc-BiN accuracy, they had greater tolerance for delayed consequences (i.e., lower discounting rate). For individuals who have higher level of Inc-BiN accuracy, rules may have had stronger stimulus control over his or her behavior, possibly resulting in greater tolerance for delayed consequences. This means that inducing Inc-BiN or improving delay discounting could lead to the emergence or improvement of the other. Past studies showed that interventions such as multiple exemplar instruction successfully induced Inc-BiN (Gilic & Greer, 2011; Greer et al., 2005), and trainings incorporating progressive delay successfully taught individuals to select LLR over SSR more frequently (Dixon & Tibbets, 2009; Dixon & Falcomata, 2004). Future studies should investigate whether inducing Inc-BiN would lead to a lower rate of delay discounting or whether teaching LLR over SSR would lead to a greater degree of Inc-BiN. This extension would inform researchers and practitioners in the design of behavior-change procedures that facilitate repertoires that better promotes tolerance for delayed consequences.
Second, the findings add to the literature on the relation between academic achievement, Inc-BiN and delay discounting. Different from the findings in this study, past studies showed that academic achievement significantly correlated with temporal discounting (Lee et al., 2012; Freeney & O’Connell, 2010). However, these studies had a large sample of adolescents as participants while the current study had a small sample size consisted of 6–8-year-old first and second graders. Younger students may not have had sufficient history of testing and the scores may not accurately reflect students’ level of academic achievement. Thus, the percentiles may not accurately reflect the younger students’ level of academic achievement. Future studies should incorporate additional dependent variables for academic achievement such as observational data and teacher devised tests or worksheets. Further, past studies showed that when individuals have Inc-BiN in repertoire, they learn at a faster rate and from exposure to instructional demonstrations in group settings (Greer & Ross; 2008; Hranchuk et al., 2019). Given that academic achievement did not relate to the accurate Inc-BiN responses, the degree of Inc-BiN may relate to the rate of learning or how fast an individual learns (e.g., fewer number of instructional opportunities) rather than academic achievement. Here, academic achievement refers to a short-term measure that captures how a student performs academically at a given moment while the rate of learning refers to a long-term measure that requires at least two time points of data to capture the "growth” or difference in achievement levels.

There were some limitations that are worthy of discussion. First, the sample size was small. Eleven students who did not demonstrate a discounting effect were removed from the analysis so there was only 13 out of 23 participants included for the final analysis. However, past delay discounting studies with adults excluded up to 15% of their data because of multiple preference reversals (Critchfield & Atteberry, 2003), and a study with sixth graders only
included 56% of the participants who demonstrated a discounting effect after applying the criteria of allowing up to two preference reversals (Reed & Martens, 2011). The current study only had 57% of participants included in the analysis, but this may be because the participants in the current study were younger compared to past studies. Children between the ages of 6 and 12 years begin to understand the value of saving and the relation between saving and more advantageous future opportunities (Te’eni-Harari, 2016). This means that the concept of saving and time may have not been developed yet for several participants in the current study, which accounts for the large proportion of the participants with variability in their delay discounting data.

Second, additional academic achievement measures would have better represented the participant’s academic performance level. The participants completed the i-Ready diagnostic in the beginning of the year where they might have experienced a “summer slide” or summer learning loss. Past studies showed that students from a lower socioeconomic background experienced an even greater learning loss over the summer due to less learning opportunities (Alexander, Entwisle, & Olson, 2007). This means that the beginning of year percentile should be interpreted with caution. One way to address this is to incorporate multiple i-Ready diagnostic percentile across the year to obtain the level of growth from the beginning to the end of the year. Future studies should also incorporate multiple academic measures including observational data, multiple testing platforms (i.e., online and paper), and teacher devised worksheets.

Despite the limitations, this investigation demonstrated a strong relation between the degree of Inc-BiN and delay discounting measures in younger children. With limited line of research evaluating the correlating variables with implications on delayed consequences, this study represents a first step in bridging the gap between the two distinct realms within behavior
analysis – verbal behavior and delay discounting. Additional research could help further shed light on promoting positive behavior change with long-term educational, behavioral, and health implications.
References


https://doi.org/10.3389/fpsyg.2014.00599


https://doi.org/10.1901.2001.76-235


American Institutes for Research.

https://charts.intensiveintervention.org/aprogressmonitoring


### Table 1

Nonparametric Correlations (Spearman’s rho) among i-Ready Scores, Inc-Bin, and Delay Discounting Measures after exclusion criteria applied (N = 13)

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. i-Ready Math Percentile</td>
<td>71.08</td>
<td>28.82</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. i-Ready Reading Percentile</td>
<td>66.77</td>
<td>23.98</td>
<td>.54</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Inc-BiN Listener</td>
<td>20.85</td>
<td>4.06</td>
<td>.26</td>
<td>.00</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Inc-BiN Speaker</td>
<td>11.31</td>
<td>4.33</td>
<td>.31</td>
<td>.43</td>
<td>.73**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Subjective Value of $100 at 1-Month Delay</td>
<td>37.32</td>
<td>32.67</td>
<td>.32</td>
<td>.20</td>
<td>.68*</td>
<td>.69**</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. AUC</td>
<td>.17</td>
<td>.22</td>
<td>-.05</td>
<td>.24</td>
<td>.47</td>
<td>.47</td>
<td>.62*</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7. k Value</td>
<td>.31</td>
<td>.49</td>
<td>-.45</td>
<td>-.18</td>
<td>-.80**</td>
<td>-.60**</td>
<td>-.90**</td>
<td>-.58*</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note. Inc-Bin = Incidental Bidirectional Naming; AUC = Area Under the Curve.*

* *p < .05, **p < .01*
Appendix

Example of the BRIC Task Datasheet (from Critchfield & Atteberry, 2003; Reed & Martens, 2011)
Chapter 6: General Discussion

Applied behavior analysis (ABA) has a strong scientific foundation of research to systematically change socially significant behavior (National Autism Center, 2015). Behavior scientists often replicate the findings from basic research conducted in controlled laboratory settings to better inform everyday practice in applied settings (Lerman, 2003). In education, educators are increasingly incorporating scientific teaching pedagogies and behavioral strategies into everyday teaching as using evidence-based practice accelerate student learning and lead to successful school improvement (U. S. Department of Education, 2016). Because educational settings often involve limited resources such as time, funds, supplies, and even human labor, it is important that these interventions and teaching strategies are scientific, effective, and efficient. As a result, “cost-effective empirically supported interventions” (Reed et al., 2013, p.37) become necessary in educational settings.

Behavioral economics principles can inform educators and researchers on the development of these scientific and cost-effective interventions for educational settings. Behavioral economics concepts are often intentionally or unintentionally incorporated into educational practices. For example, teachers provide praise and small goods to reinforce students’ appropriate behavior such as assignment completion and sitting nicely. Response and reinforcement are constantly in play and manipulated by educators to advance learning. Response-reinforcer relations are often investigated under the realm of behavioral economics, suggesting the feasibility of the practical application of behavioral economics principles within an educational setting. Thus, more studies on applying behavioral economics to educational settings are needed (Reed et al., 2013; Staubitz et al., 2018).
The four studies in this paper were direct attempts to fill that gap. The systematic review and meta-analysis in Study 1 bridges the two seemingly distinct fields – behavioral economics and education. Token economy is a prominently used evidence-based classroom management strategy (Simonsen et al., 2008) with its conceptual foundation on response-reinforcer relations under behavioral economics. Study 1 identified 24 token economy studies conducted in kindergarten to 5th grade settings and compared the token economy components within general and special educational settings to understand the components and response-reinforcer arrangements contributing to behavior change. The findings showed that token economy interventions are highly effective in changing behavior with large effect sizes, regardless of the type of classroom. Also, token economy components including backup reinforcer types, token production rate, and exchange production rate differed between the classroom types.

The findings of the systematic review and meta-analysis suggested future directions for practitioners in terms of response-reinforcer arrangement in a classroom setting. When analyzing the effects of a token economy, it would be important to consider the embedded schedules of reinforcement within the token economy system. Token economies inherent three interconnected schedules of reinforcement including token production (i.e., how often tokens are delivered, or the schedule of reinforcement for token), exchange production (i.e., how often tokens are exchanged, or schedule of reinforcement for backup reinforcers), and token exchange (i.e., the cost of backup reinforcers in tokens) schedules (Hackenberg, 2009; Ivy et al., 2017). These components are essential in making a token economy system flexible and adaptable to various settings and populations. With the findings in the review, practitioners can arrange the token production and exchange production rate most effectively for their respective classroom type. For example, the token production rate could be higher for students in special education settings.
to increase responding and to build in greater opportunities for the students to contact reinforcement. Overall, in 2000-2019, there was only a handful of high-quality token economy studies conducted in educational settings despite token economy being prominently used in classrooms. There is a greater need for high quality token economy studies conducted in educational settings to improve fidelity in replications for future research and practice. Future researchers could use the What Works Clearinghouse (WWC) standards for single-case design (Kratochwill et al., 2010) to ensure that the studies adhere to the minimal standards for high-quality single-case design studies. One limitation worth mentioning is the lack of proper publication bias measure. Future studies should incorporate an appropriate publication bias measure as indicated in the PRISMA guidelines. For example, future researchers could compare unpublished dissertations to published paper effect sizes or compare low quality study effect sizes to high quality study effect sizes.

As a next step, it was important to investigate how token economies functioned in an applied setting and whether it related to socially significant behaviors. The goal of Study 2 was to utilize a classwide token economy system in a general education setting to (1) develop a rapid, user-friendly assessment for sensitivity to delayed rewards and (2) determine the relation between students’ sensitivity to delayed rewards and saving and spending behavior in a token economy system. The two experiments collectively showed that the 1-month delay indifference point predicted classroom behavior and that students who discounted less and had greater self-regulation accrued and saved more tokens. Further, there were group differences in how students respond to token production schedule change and a variable token production schedule better correlated with discounting compared to a fixed schedule.
A brief, user-friendly assessment has important implications in an educational setting. Educators often need to evaluate students’ level of self-regulation (often referred to as “self-control” or “impulsivity”) and add the information in report cards as part of an evaluation of student behavior in class. A commonly used behavioral assessment to measure individuals’ level of sensitivity to delay is delay discounting. However, the complex nature of the assessment procedure and deriving meaningful numbers impede its use among educators in their daily practice (Critchfield & Reed, 2009). Thus, it is crucial to develop a more accessible assessment for educators. The BRIC Task offers a simple assessment method by utilizing just one delay at 1-month instead of multiple delays used in the traditional delay discounting assessments. The BRIC Task is not by any means a substitute of the traditional delay discounting assessment, but rather an alternative offering a quick, valid measure of sensitivity to delayed rewards that educators or practitioners can use without deriving additional data using a software or mathematical equations. Using the data, educators can arrange the classroom contingencies to increase appropriate student responses. For example, if a student scores higher on the BRIC Task (i.e., higher level of self-regulation), educators could use a leaner token production rate which requires greater number of responses to earn a reinforcer. On the other hand, if a student scores lower on the BRIC Task (i.e., lower level of self-regulation), educators could use a richer token production rate which requires smaller number of responses to earn a reinforcer. For this student, the teacher could provide more frequent reinforcement such as praise and tokens when learning new skills to increase the target behavior.

Another important implication of the findings is the development of an intervention increasing tolerance to delays or, in other words, promoting advantageous choice-making behavior in younger children. Past researchers have used progressive delay to successfully teach
individuals to select a larger, later reinforcer over a smaller, sooner reinforcer across individuals with developmental disabilities (Vollmer et al., 1999), emotional and behavioral disorders (Staubitz et al., 2020), brain injury (Dixon, & Tibbetts, 2009), and ADHD (Binder et al., 2000). However, these studies were all done on a one-to-one basis where the experimenter delivered the intervention one individual at a time. In a general education setting, most instructions are delivered to large or small groups and one-to-one instruction may not be feasible when there is only one teacher with multiple students in the classroom. Thus, it was worth investigating a group-based intervention using progressive delay to improve advantageous choice-making behavior of students in the classroom.

Study 3 directly addressed the implication for future research in Study 2. The findings suggested that that students’ sensitivity to delayed rewards was related to classroom behavior and that there is a need for a group-based intervention promoting advantageous choice-making behavior. Study 3 aimed to improve second-grade students’ advantageous choice-making behavior in a general education setting by introducing a group-based intervention involving progressive delay and choice opportunities. The results showed mixed results – students performing on and below grade-level in math showed increased levels of advantageous choice making behavior while the same effects were not shown in students performing above grade-level in math.

The discrepancy of the results could be explained in three possible ways. First, students’ level of self-regulation, which was measured using the BRIC Task, prior to the intervention may have contributed to the differential effects. Several studies have shown that having strong self-regulation skills is related to school success in both math and reading (Adams & Snowling, 2001; Connor et al., 2010; Day et al., 2015; St. Clair-Thompson & Gathercole, 2006). This means that
students performing above grade-level in math already demonstrated a high level of self-regulation and thus, there might have been a ceiling effect at increasing advantageous choice-making behavior. Second, the intervention may have functioned as punishment for students functioning at above grade-level in math. The above grade-level students were already selecting the larger, later reinforcer prior to the intervention and experiencing the intervention may have been aversive. This speaks to the importance of educators considering each student’s academic level prior to the implementation of the intervention. Third, students’ levels of verbal behavior could explain the discrepancy. Skinner (1957) argued that delayed consequences are often controlled by verbal statements of a contingency or rules, which are contingency-specifying stimuli produced by one’s own verbal behavior (Zettle, 1990). To have delayed consequences reinforce behavior, an individual must be able to hear oneself say the rules with specific contingencies. Students who performed above grade-level in math often performed above grade-level in reading and writing, indicating that they functioned as readers and writers (Greer & Ross, 2008). If these students already functioned as speaker-as-own-listeners, they would be able to recite the rules to themselves to modify their own behavior and thus demonstrate higher level of advantageous choice-making behavior.

Further, the findings also contribute to the existing literature on ADHD and delay discounting. Literature on children with ADHD showed that younger children with ADHD prefer more immediate rewards compared to younger children without ADHD (Paloyelis et al., 2009; Tripp & Alsop, 2001). Because the current study showed that it is possible to improve advantageous choice-making behavior using the progressive delay training, for children with ADHD, delay discounting or the BRIC Task can be used to match intervention. Practitioners can
utilize more immediate reinforcers of a greater magnitude to change behavior, highlighting the utility of delay discounting in educational settings.

Overall, the findings in Study 3 were consistent with past studies that successfully demonstrated increased level of advantageous choice-making behavior by gradually increasing delays to the larger reinforcement with choice opportunities (Fisher et al., 1997; Mazur & Logue, 1978; Schweitzer & Sulzer-Azaroff, 1988). The results also supported past research that showed increased advantageous choice-making behavior when individuals were provided a concurrent activity during the delay (Dixon & Cummings, 2001; Dixon & Holcomb, 2000). In the context of the classroom, the findings suggested that educators should incorporate progressive delay and choice-making opportunities to instruction or token reinforcement system to provide increased opportunities to teach advantageous choice-making behavior throughout the day.

Given that different levels of verbal behavior may affect students’ advantageous choice-making behavior, the goal of Study 4 was to investigate the relation between classroom variables, including verbal behavior, contributing to tolerance to delayed consequences. Thus, Study 4 investigated the relation between verbal behavior repertoires, delay discounting, and academic achievement. The results showed that the accuracy of Inc-BIN positively related with delay discounting measures including the $k$ value and subjective value of $100$ at 1-month delay obtained using the BRIC Task. On the other hand, there was no significant relation between academic achievement and either delay discounting or Inc-BiN.

Study 4 was the first direct attempt at bridging the fields of verbal behavior and behavioral economics. The findings suggested that verbal behavior and behavioral economics are not mutually exclusive and can inform one another to develop innovative practices for educational settings. For example, a behavior-change strategy stemmed from a collaborative
effort between the two fields could include an intervention teaching a child to verbally recite the delay and surrounding contingencies to oneself while waiting to contact the delayed reinforcer. It may be necessary that children can emit speaker and listener responses for delayed stimuli to function as reinforcement.

The four studies, individually and collectively, contributed to understanding how behavioral economics can be applied to education. Collectively, the studies showed that token economies relate to basic behavioral economics principles and that we could use token economies in a progressive manner to teach students to wait for the larger, later reinforcer. When using token economies, educators could incorporate progressive delay across multiple academic subjects and non-academic subjects to promote advantageous choice-making behavior. An important point is that Study 2 and 3 manipulated the token production rate by progressively delaying the delivery of tokens. However, Study 1 clearly showed that the other reinforcement schedules, exchange production and token exchange rate, also affect student responding and effectiveness of the overall token economy system in educational settings. It would be beneficial to investigate how the manipulation of each token economy reinforcement schedule affects students’ advantageous choice-making behavior. Past basic research (Bullock & Hackenberg, 2006; Foster et al., 2001) showed that the token production and exchange schedules affect each other because the first schedule of reinforcement is reinforced by the second schedule. On the other hand, applied research showed that the changes in exchange schedules in the token economy did not change responding toward the token production schedule (Argueta et al., 2019). To better understand how token economy schedules interact with one another in an applied setting, perhaps, a comparative analysis that directly compares the effects of the manipulation of
each token economy reinforcement schedule on student responding would be beneficial. The findings would inform educators on a more effective progressive delay procedure.

Additionally, educators should focus on inducing missing verbal behavior cusps while using token economies to change relevant behavior. Specifically, speaker-as-own-listener cusp, including Inc-BIN, should be taught for one to respond as a listener and provide for their own speaker behavior (Greer & Ross, 2008; Greer & Keohane, 2006). There have been several interventions used to induce Inc-BIN such as multiple exemplar instruction (Gilic & Greer, 2011; Greer et al., 2005). Token economy systems are often incorporated into daily instruction and during these interventions. This means that during the interventions, educators and practitioners can use token economies at the same time as delivering instruction to induce the cusp and to change socially significant behavior. Thus, future researchers should investigate whether incorporating progressive delay of token delivery to Inc-BiN intervention procedure can teach both Inc-BiN and advantageous choice-making behavior. Similarly, future studies should determine whether inducing Inc-BiN have a similar effect on improving advantageous choice-making behavior as the progressive delay procedure did. This procedure would be efficient and beneficial for the students because the behavior would come under the control of natural contingencies, which would in turn facilitate maintenance of the behavior. It also may be beneficial for future researchers to report the token economy schedules that were used during the intervention sessions so future researchers can better replicate the procedure.

The studies also collectively suggest the effects of conditioned reinforcement for academic content on tolerance to delayed consequences. Past studies showed that conditioned reinforcement for the observation of stimuli increases the reinforcement value and results in faster and immediate learning (Gentilini & Greer, 2021). Specifically, these studies investigated
the effects of conditioned reinforcement for reading (Bly & Greer, 2019; Gentilini & Greer 2021), three-dimensional stimuli (Du et al., 20), and adult faces and/or voices (Maffel et al., 2014). In Study 3, progressive delay training procedure did not increase the percentage of LLR selection for participants performing above grade-level in math. Conditioned reinforcement for math may have been in repertoire for these participants given that the participants enjoyed solving math problems or, in other words, allocated their behavior in engaging in solving additional math problems instead of solving zero problems. Participants who enjoy math may demonstrate greater tolerance to delayed consequences involving math problems. Future researchers should study the relationship between tolerance to delayed consequences and conditioned reinforcement for academic content including math, reading, and writing, and develop interventions that directly target the academic content functioning as conditioned reinforcement to improve advantageous choice-making behavior.

Future studies should also investigate whether conditioned reinforcement for peers affect student responding. The meta-analysis (Study 1) showed that token economies for students in general education classrooms involved more peers and social competition to increase responding. Further, the researchers reported in Study 2 and 3 the potential effect of peers on the participants’ responding to token economies and progressive delay training. These studies suggest the role of peers on students’ advantageous choice-making behavior. When peers function as conditioned reinforcement, greater conversational units between a student and his or her peer may occur and in turn, this student may alter his or her own behavior based on the peer’s behavior. If this student and the peer discuss saving the tokens or selecting the LLR over SSR, this student would more frequently save tokens or select the LLR over SSR. Therefore, it may be worth
investigating the relationship between the degree of conditioned reinforcement for peers and advantageous choice-making behavior.

6.1 Summary

Overall, the findings of the four studies are important for both practice and research for two main reasons. First, the data support that behavioral economics concepts inform educators on efficient and effective instructional strategies and assessments. Study 2 and 3 provided an assessment and intervention that could effectively and efficiently change or measure socially significant behavior that are of interest to educators. Second, the data support the conceptual intersection between behavioral economics and education. Study 1 and 4 empirically supported the connection between the two seemingly distinct fields by showing that token economies relate to basic behavioral economics principles and that we could use token economies with progressive delay to teach students to wait for more advantageous, delayed reinforcer. However, these studies only serve as a foundation to connect the two fields. Future researchers should continue to strive to investigate the conceptual framework that explains the connection and develop efficient and effective instructional strategies based on the framework. Given the clear intersection between the behavioral economics and education, the field of ABA would benefit from more empirical studies that shed light on the complex relation between the two fields and extend the findings for the development of innovative teaching strategies and technologies.
References


