5-2021

The Effects of Token Menu Manipulations on Token Demand

Sean Regnier
Western Michigan University, seanregnier9726@gmail.com

THE EFFECTS OF TOKEN MENU MANIPULATIONS ON TOKEN DEMAND

by

Sean Regnier

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Psychology
Western Michigan University
May 2021

Doctoral Committee:

Anthony DeFulio, Ph.D.
Wayne Fuqua, Ph.D.
Jon Baker, Ph.D.
Tim Hackenberg, Ph.D.

THE EFFECTS OF TOKEN MENU MANIPULATIONS ON TOKEN DEMAND

Sean Regnier, Ph.D.

Western Michigan University, 2021

Token economies are systems of contingencies that are designed to reinforce targeted behavior. Engaging in a targeted behavior produces conditioned stimuli that can later be exchanged for established reinforcers. A back-up reinforcer is an established reinforcer that can be acquired by exchanging the tokens. A component of token economies that has received little attention in the literature is the composition of the set of back-up reinforcers available for exchange; typically referred to as the menu. When used as part of behavior therapy, the token menu often contains a set of items that has been identified by conducting a preference assessment or interview. Absent an empirical basis for doing otherwise, most decisions about the structure of the menu are made for logistical reasons, and in some cases may reflect nothing more than what is convenient for the therapist. The overall purpose of the present studies was to determine the effects of token menu manipulations and token component schedules on demand for tokens. In the main study, the token production schedule, types of items on the menu, and the number of items on the menu were manipulated to assess their effects on demand for tokens. Essential value increased as the number of items on the menu increased for all four study conditions. A statistically significant interaction was observed. This interaction involved the effects of the number of menu items and reinforcer category on essential value. Specifically, the mixed 12-item menu produced the highest essential values, and the primary reinforcer 3-item menu produced the lowest essential

values. From a translational perspective, these results highlight the importance of including back-up reinforcers that are related to a variety of motivational operations whenever possible when designing token economies, and that larger menus should outperform smaller menus, at least across the range of values used in the present studies. Future research should involve the implementation of a mixed model with more narrow constraints on response parameters to control for variability in the dataset.

ACKNOWLEDGEMENTS

There is no higher level of praise and appreciation that I can provide for Anthony DeFulio, my mentor and supervisor for the past four years. He has picked me up when I felt down; pushed me when I was overly content; and put me in a position to succeed. He is a large reason why I remain humble yet carry myself with a level of confidence needed to succeed in this field. He remained appreciative of my multiple emails per day, ranging from questions about current research projects, emotional reactions when working with our research participants, to some of my insane research ideas around the stock market. However, by far the most significant impact he has had on me was by accidently introducing me to my follow lab-mate and now fiancé, Haily Traxler, whoops.

Haily, there isn't enough room for me to list my appreciations for all you have done to help my on my journey over the past four years. Moving to Kalamazoo on my own with two cats was much more difficult than I had expected. There were times where I had serious doubts of my ability to complete the program. Those worries went away with you by my side. Cheers to the next chapter!

I would also like to thank my friends and family back in Massachusetts. When I first moved to Kalamazoo, I didn't consider the burden that would have on my family, especially my parents and young nieces and nephews. They have remained supportive despite me not being home in close to a year. Their frequent Zoom and phone calls providing reassurance helped get me through the times I was especially homesick. I have kept a close group of friends for the past

25 years. Together we have gone through several tragedies in the past few years. They kept me motivated and gave my work around substance abuse and mental health further purpose.

Thank you to the lab for providing me relief via walkies and beeries and helping mold my research ideas. Mark, I appreciate you pushing me to continue through with my statistics certificate and helping me figure out MTurk. The latter is still a process.

I also want to thank my dissertation committee with all their support over the past year. You were large contributors to my thesis replacement, my comprehensive exam, and now my dissertation. A lot of hard work goes into reviewing these materials, and I do not take that for granted.

Finally, I wish to thank my previous mentors Dr. Joey Reyes, Dr. Leonardo Andrade, Dr. Cynthia Anderson, and Meg Walsh. Each of you pushed me to go beyond the work of a typical BCBA. I wouldn't have even applied to Western if it weren't for your encouragement. You clearly saw something in me that I didn't at the time. I will be forever in your debt.

Sean Regnier

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Conditioned Reinforcement

Conditioned reinforcement is a behavioral process in which a previously neutral stimulus acquires value due to its association with a primary reinforcer or another conditioned reinforcer (Williams, 1994). Because condition reinforcement can be used to account for the development and maintenance of behaviors that are not directly related to survival, it is thought to have broad applicability to human affairs. In laboratory experiments, it is typical for a conditioned reinforcer to be correlated with a single, primary reinforcer (e.g. food or water) that is established by a singular deprivation operation. Although simple arrangements are often preferred in laboratory settings, there is no theoretical or conceptual requirement that restricts the number of backup reinforcers that can be correlated with a conditioned reinforcer.

Generalized Conditioned Reinforcement

A conditioned reinforcer that is paired with multiple primary reinforcers is known as a generalized conditioned reinforcer. According to Skinner (1953) a generalized conditioned reinforcer may exert more control over responding because they can maintain behavior under multiple states of deprivation. For example, lever presses and key pecks reinforced with a clicking sound that have previously been paired only with food would eventually decrease over the course of a session as a result of food satiation. However, if those clicks had been paired with water as well as food then subjects would likely respond longer as responses are under the control of two deprivation operations.

One of the most common uses of generalized condition reinforcement as a part of everyday life and clinical interventions, is the use of social reinforcers such as praise and

approval. Interventions using social reinforcers have been widely published for over fifty years (Allen et al., 1964; Zimmerman & Zimmerman, 1962) and continue to be a common component of behavior treatment plans for individuals with developmental disabilities and mental illnesses. The efficacy of social recognition as a conditioned reinforcer extends to the field of Organizational Behavior Management (Stajkovic & Luthans, 1997; 2006) where it is typically correlated with the presentation of other conditioned reinforcers, such as money or promotions to improve performance in the workplace (e.g. Crowell et al., 1988).

There are several other types of conditioned reinforcers that have been used to promote skill acquisition or reduce problematic behavior. For example, Ferster and DeMyer (1962) used several creative coin operated devices to train matching techniques to children with developmental disabilities. Coins could be deposited into several machines that would activate, such as pinball, vending machines, a pigeon and monkey that would perform when the inserted coin illuminated a light, a television, electric train, etc. The authors found that they were reliably able to bring responding under the control of artificial conditioned reinforcers and potentially widen their repertoire further by manipulating the conditions under which coins were earned or machines were activated.

One of the most prevalent generalized conditioned reinforcers is currency. Money is effectively established as a reinforcer by an enormous number of motivating operations in that it can be exchanged for anything that can be bought (Skinner, 1953; Kelleher & Gollub, 1962). Financial incentives are a component of everyday life, ranging from a paid salary to the most robust clinical intervention for substance abuse through contingency management (Davis et al., 2016). Contingency management (CM) is an intervention in which material incentives, most commonly money, are delivered to participants contingent on evidence that they have engaged in

a specific target behavior (Higgins & Silverman, 1999). One of the reasons why CM interventions are so effective is the generalized conditioned reinforcing properties of money which can compete with the most powerful primary reinforcers, such as drugs like heroin.

In addition to the independence of deprivation conditions, there are several other advantages to using generalized conditioned reinforcers (Kazdin & Bootzin, 1972). Generalized reinforcers can maintain performance when the terminal reinforcer is temporarily unavailable and can maintain target behavior in the face of changes in preferences for back-up reinforcers, whether temporary or permanent (Skinner, 1953). Additionally, generalized reinforcers have higher reinforcing value. While this statement was made by Skinner in *Science and Human Behavior* and is possibly the most significant advantage of generalized conditioned reinforcers, the empirical evidence supporting his claim is limited (see DeFulio et al., 2014; Traxler & DeFulio, In Prep).

Token Economies

The use of generalized conditioned reinforcers over typical conditioned reinforcement procedures provides robust clinical utility. When these reinforcers are physically manipulable objects (e.g., poker chips, coins, bingo chips, etc.) they can have further advantages over other generalized conditioned reinforcers (Kazdin & Bootzin, 1972). An example of a tangible generalized conditioned reinforcer is the token economy.

Token procedures involve reinforcing desirable behavior with tokens that can later be exchanged for back-up reinforcers. A back-up reinforcer is anything that can be acquired by exchanging the tokens. Back-up reinforcers are also called terminal reinforcers because they are the outcome of the component schedules of token reinforcement. Like other generalized conditioned reinforcers discussed previously, token economies offer advantages relative to

providing the back-up reinforcer directly without the use of tokens (Kazdin & Bootzin, 1972; Kazdin, 1982). One of the most important advantages is that token reinforcers can bridge the delay between a response and a terminal reinforcer (Kelleher, 1966; Skinner, 1953; Wolf, 1936). For example, in the second order schedule research by Kelleher (1966) rats' lever presses were maintained for one hour without access to primary reinforcement. Incentives provided in CM interventions can be saved rather than spent immediately (Subramaniam et al., 2017) without losing their value as the delay to spending increases. However, the tangibility of tokens imparts an advantage over other generalized conditioned reinforcers, such as social reinforcers (Ayllon and Azrin, 1968a; Kazdin & Bootzin, 1972). Specifically, tokens provide an objective, quantifiable measure of reinforcer magnitude (Ayllon & Azrin 1968). The number of tokens an individual receives corresponds to the amount of reinforcement provided. Additionally, tokens can remain in an individual's possession outside of the context they are earned and spent, can be accumulated without limit, and may maintain their value over time.

Token economies have additional advantages that increase their practical value. For example, in situations like those described in Ferster and DeMyer (1962) back-up reinforcers can be delivered automatically, such as a vending machine. Additionally, practitioners can standardize the physical characteristics of tokens to be unique to an individual's preferences. Finally, tokens are easily paired with other sources of sources of reinforcement, such as social praise or approval. In sum, tokens have broad utility in laboratory and applied settings and are the most potent and practical method for reinforcing behavior in clinical and educational behavior modification procedures.

Applications of Token Reinforcement

The clinical application of token economies began in the 1960s with the seminal work of Allyon and Azrin (1965). In this study a token economy was used to increase socially significant behaviors with individuals with mental illnesses who lived in a psychiatric facility. Tokens were presented contingent upon engaging in useful behaviors in the facility, examples including serving meals, cleaning, clerical work, and sorting laundry. They could be exchanged for a wide variety of back-up reinforcers such as leaving the facility, social interactions with preferred individuals, recreational activities, toiletries, decorative items, and food. According to Allyon and Azrin (1965) tokens were specifically used to bridge the gap between the target behavior and terminal reinforcer and was inspired by the second-order schedule research of Kelleher (1957). Tokens effectively improved and maintained performance, which decreased when the token contingency was removed. Additionally, the tokens provided an objective measure of performance and reinforcement delivery.

Allyon and Azrin's work served as a catalyst for the generalization of token economy research to other populations. Some examples over the past fifty years include using token economies to increase bus taking in crowded cities (Deslauriers & Everett, 1971), increase attendance and task completion with children in juvenile court (Phillips et al., 1971), promote independence and pro-social skills for individuals with mental health disorders (Paul & Lentz, 1977), reduce illicit drug use (Glosser, 1983), and promote classroom participation (Boniecki & Moore, 2003). Token economies are easily disseminated and robust interventions that have been implemented in dozens of settings with many different types of individuals.

Translational Research in Token Reinforcement

Experimental and clinical applications of token economies were prominent for three decades following the work of Allyon and Azrin in the 1960s. However, this research has generally decreased over the last twenty years even though the translational value of token economies is profound. In addition to providing insight on the ability of token economies to promote healthy behavior with a variety of clinical populations, token economy research has significance that reaches to the experimental analysis of behavior, behavioral economics, and behavioral ecology. Through laboratory studies, token economies have provided important information about basic behavioral processes. Wolf (1936) and Cowles (1937) used tokens as some of the earliest demonstrations of conditioned reinforcement. Studies using token economies (e.g. Traxler & DeFulio, In Prep; DeFulio et al., 2014; Tan & Hackenberg, 2015; Andrade & Hackenberg, 2017) have tested the Skinner's (1953) claims of the reinforcing value of generalized conditioned reinforcers.

It is important to note that currency is by definition a token in that it is a physical, generalized conditioned reinforcer that can be exchanged for other reinforcers. Therefore, token economies serve as reasonable analogs for monetary spending in controlled settings. Token economy research also allows for breaking down cost into components and studying their separate effects on responding. This research is thus valuable for the field of behavioral economics. For example, in one of the earliest collaborations between economists and psychologists, Battalio et al. (1974) used a token economy to measure consumer behavior in a psychiatric facility. The authors measured several variables, primarily price which was manipulated by changing the cost of the back-up reinforcers in the facility's marketplace. Price can be described as the token exchange schedule of a commodity (Hackenberg, 2018). They

were then able to assess how the consumption of various commodities was a product of cost. Token economies also allow for the controlled manipulation of wages via token production schedules by manipulating the amount of work someone must do receive payment.

The final contribution of token economies extends to the field of behavioral ecology (Cronk, 1991; Winterhalder & Smith, 2000). Human behavioral ecology involves the application of models of evolutionary ecology to human behavior (Williams & Fantino, 1994). As summarized by Cronk, the emphasis of human behavioral ecology is on the relationship between evolution by natural selection and the behavior of modern humans. Token accumulation research can contribute to current ecological models of foraging. For example, the marginal value theorem (Charnov, 1976) models the relationship between travel cost and time spent in patches. In the context of foraging, animals typically find food in patches. According to the Optimal Foraging Theory, predators forage in a way that maximizes their fitness (Williams & Fantino, 1994). Therefore, careful consideration must be made of travel distances from one patch to another, and travel time spend at each patch. Token economies serve as an excellent representation of foraging models because exchange production schedules can act as travel costs, and accumulation can represent time spent in patches.

As a generalized conditioned reinforcement procedure, token economies are important in the research comparing the optimal foraging theory and the delay reduction hypothesis, discussed earlier in the manuscript. According to the optimal foraging theory predators rely on the maximization of reinforcement. or. According to Williams and Fantino however (1994), predators primarily direct their responding towards stimuli that signal a greater reduction in waiting time (Williams & Fantino, 1994). This is understandable given what is known about reinforcer value. Generally, the value of a reinforcer (V) is determined by its magnitude (A) and

the delay to receiving it (D; Mazur, 1987). Value can be calculated using the equation V= A/(1+kD). An inverse relationship between value and delay can be observed. In other words, increasing delays to a reinforcing stimulus would decrease its reinforcing value, which is why an organism would search for signals of a reduction in delay. In terms of token accumulation however, as an individual accumulates tokens, they are increasing the delay to making the exchange. This would suggest that accumulation decreases the value of a reinforcer, and yet accumulation occurs in most applied and laboratory token economy research. This may indicate that delay reduction hypothesis require flexibility in the face of generalized conditioned reinforcers.

Efficacy of Token Reinforcement

The token economy is one of the most widely used treatments in behavior analysis. This intervention has its own place in the Behavior Analytic Certification Board's (BACB) fifth edition task list (see G-17 Use of Token Economies). They have been successfully implemented with countless types of individuals and behaviors. In a meta-analysis of token economy research in classrooms from 1980-2014, Soares et al. (2016) found a weighted effect size of .82, indicating that token economies were strongly effective interventions at reducing problem behavior and improving academic skills, especially when used in classrooms. While using target behavior reduction as an index of token utility can provide an overview of their efficacy, it is difficult to assess the contributions of smaller token components without a more thorough analysis.

When token economies are implemented in an applied setting, it is common for designers to emphasize the amount of work an individual must do to earn a token. However, token economies contain several components from the moment a token is earned to the time it is

exchanged for a back-up reinforcer. These can be described as component schedules and each can have differential effects on behavior when manipulated. There are three component schedules in a token system: the token production schedule, the exchange production schedule, and the token exchange schedule (Hackenberg, 2009; 2018). Before defining the three component schedules it is important to note that token economies can be arranged as second order schedules (Kelleher, 1966). Second order schedules occur when the completion of responses on one schedule serves as a unit of another schedule. In the context of a token economy earning a token and exchanging it operate on separate schedules. If a participant must emit 10 responses to earn a token but can only exchange their tokens after earning five tokens, their responding is operating under a second-order schedule. Therefore, the three component schedules are components of a second-order schedule.

The token production schedule describes the contingency in which a token is earned. For example, an individual may need to complete a daily living activity to earn a token while a rat may need to press a lever five times to produce a token. Using prior examples, in Kelleher (1957) chimpanzees' responses were operating under an FI-5-minute token production schedule with the first response after five minutes producing a token. The exchange production schedule describes the conditions under which an exchange period is produced. In practice, people participating in a token reinforcement procedure are often required to earn a specific number of tokens or to earn tokens for a specified time period prior to having the opportunity to exchange. Continuing with the Kelleher (1975) example, the chimpanzees completed six consecutive FI schedules to exchange their tokens for food. Therefore, responding was operating under a FI-5-minute *token* production schedule and a FR-5 *exchange* production schedule.

Token exchange schedule describes a response that is required to exchange tokens for back-up reinforcers. After an exchange schedule has been completed, a pigeon may need peck a key to produce 2-seconds of food access (Yankelevitz et al., 2008). Token exchange schedules are typically held constant in laboratory settings in order to assess the differential effects of the other component schedules. In an applied setting, the token exchange schedule may include a manipulation of price after the exchange production schedule has been completed by the participant.

Common Token Economy Procedures

All token economies follow a basic format: an individual receives a token for meeting a behavioral requirement. At some point, that individual can exchange their tokens for a back-up reinforcer of their choosing. While this serves as the foundation for a token economy procedure, several variations are common. One procedure that is common in applied settings is the token board procedure. Under this procedure, participants are typically required to earn a certain number of tokens to produce an exchange opportunity. For example, a student may earn one token for sitting in their seat for 2 minutes during a table activity. Each token earned is placed on a token board with five spaces. When they have filled up their board, they can exchange their tokens for the back-up reinforcer. The major difference when using a token board is that by adding the five-token requirement, the clinician has introduced a second-order schedule of reinforcement. Meeting the two-minute seating requirement is one unit in a five-unit requirement. In this example, the token production schedule was the two-minute requirement, and the exchange production schedule is an FR-5. An additional trait of token boards is that the back-up reinforcers do not typically have a token exchange schedule beyond the participant stating that they would like to exchange their tokens (FR-1). After the TP and EP schedule

requirements have been met, there is no additional response requirement. It is common for all items on the menu to have the same price, typically the cost of filling up their token board. Using the same example, after the five tokens have been earned, the student may get to exchange them for a snack or fifteen minutes of an activity like coloring with crayons or access to an iPAD.. After the access time has expired, the tokens are removed from the board and the process can be repeated. Token boards are popular because response requirement can be manipulated in several ways by increasing the requirement to earn tokens and adding or removing available spaces on the board. This allows the token economy to be gradually faded in a way that best promotes maintenance of therapeutic outcomes.

A second popular token economy procedure involves an open market. This is the closest analogue to the earning and spending that occurs with everyday purchases. In an open market, participants earn their tokens as described above. However, they are typically allowed to spend them at any time, with no additional exchange production schedule requirement. In addition to the lack of a programed exchange production schedule, this type of token economy includes a more thorough token exchange schedule requirement. This involves the manipulation of the prices of menu items. Modifying the previous example, the student still receives one token for sitting in their seat for two minutes during a table activity. However, instead of a token board, they are now presented a menu of back-up reinforcers with several options at different prices. Fifteen minutes on an iPAD may cost five tokens, coloring costs three tokens, and a snack costs 10 tokens. The token production and token exchange schedules can be tailored to specific applications or clients.

While the two types of token economies are common as described, they can both be modified to fit the needs of the individual receiving treatment. For example, a token exchange

schedule could be incorporated into the token board by adding prices to the items on the menu. In this situation the participant still must earn five tokens to exchange them, but can choose to exchange all, none, or some of the five tokens earned. When the tokens are exchanged, only the spent ones are removed from the board. The open market can also introduce an exchange production schedule by adding a rule that participants must earn a certain number of tokens before being allowed to exchange them. There could also be a travel cost, such as the market being in another classroom that participants must travel to in order to exchange their tokens. While both token economies may appear dissimilar, they are simply different combinations of the three component schedules.

The Token Menu

While token economies have produced robust treatment outcomes for decades, many of their significant components are often overlooked, or not described in the literature. For example, in a review of procedural descriptions of important token economy components Ivy et al. (2017) found that only 19% of research articles from 2000-2015 included complete descriptions of all necessary components. A component of token economies that has received little attention in the literature is the collection of possible back-up reinforcers available for exchange; typically referred as the menu. Also called the marketplace, the token menu often contains a random selection of various primary and conditioned reinforcers preferred by the participant. Examples include time playing various games or enjoying hobbies, preferred snacks, and special privileges in school or psychiatric settings (Phillips et al., 1971; Ayllon & Azrin, 1965), or monetary reinforcers in contingency management interventions (Dallery et al. 2015; DeFulio & Silverman, 2012). While it is common for clinicians to select menu items based on the preferences of their clients, there are several other menu variables that are often overlooked.

The first obvious, yet significant menu quality is the number of items included on the menu. Token economies can be arranged such that a single pre-determined item, such as a preferred snack, is delivered to a participant after filling a token board. If performance decreases, the token economy developer may change the item on the menu. Participants typically have access to the menu before they start earning tokens and can select the back-up reinforcer prior to completing a token production schedule or can make the selection after the schedule has been completed. One possible advantage of increasing menu size is that a larger selection of backup reinforcers can maintain behavior under several conditions of deprivation. A larger menu increases the generality of tokens, which should increase token value, at least across the lower range of the parameter space. However, increasing the number of items on the menu will not control responding under multiple deprivation conditions if the reinforcing value of all items is determined by the same motivating operation. For example, having three, five, or ten different brands of salty chips will not likely produce differential, radical, changes in responding. Therefore, the menu size is not very important on its own. Its effect on responding is likely moderated by the availability of different types of items on the menu.

Reinforcers used in applied settings are typically categorized by their physical properties: 1) primary reinforcers (edibles); 2) activities; 3) social reinforcers; and 4) tangibles. These categories are common in indirect preference assessment tools like the *Child Reinforcement Survey* (Fantuzzo et al., 1991) and functional assessments like the *Functional Assessment Interview* (O'Neill et al., 1997). They also have some correspondence with traditional functional analyses conditions (e.g. Iwata et al., 1994). The same reinforcers are often used as back-up reinforcers in token economies. Primary reinforcers are very common back-up reinforcers in token systems because they are useful for individuals with limited preferences. They are often

used in token economies for individuals with developmental disabilities (see Becraft & Rolider, 2015; Kazdin & Bootzin, 1972). They can also be quickly consumed with limited delays in teaching trials. Primary reinforcers in token systems can be any variety of food and drink items that are preferred by the participant. The second category of menu items are preferred activities. This could be any amount or duration of an activity which throughout history have included watching TV, outings (Winkler, 1973), special privileges, time unsupervised, access to religious services, and listing to music (Ayllon & Azrin, 1965). More modern token interventions have used similar back-up reinforcers such as arts and crafts, outings with preferred staff, and shopping outings (Nastasi et al., 2020). Activities often blend with other categories if they are enjoyed with other individuals or involve some tangible item. The next category is social reinforcers. This may include time with preferred individuals, engaging in the above activities with someone else, or just having physical contact, such as hugs or tickles. Finally, tokens can be exchanged for a certain amount of a tangible reinforcer such as time on an electronic, or purchasing miscellaneous items like DVDs, hair dye, coffee cups, and baseball cards. Cigarettes have also been a popular back-up reinforcer used throughout the history of token economies (see Winkler, 1973). It may be beneficial to have a token menu that contains multiple categories of back-up reinforcers. Increasing number of available categories would increase token generality, therefore increasing token value. Also, categories that blend (e.g. social + activities) would likely increase value as well due controlling responding under several deprivation conditions.

The number and type of back-up reinforcers in a token menu may interact with each other by affecting generality. Generality is likely one of the most powerful menu manipulations as it plays a role in the effects of several other menu variables. By having several back-up reinforcers of various types, a participant's responding can be controlled by the presence of many

motivating operations. For example, having three salty snack items on the menu may not be as effective as having one salty snack, one preferred activity, and one preferred tangible item. Having 10 salty snack items on a menu is larger than only money on the menu but is not nearly as effective for most people. Traxler & DeFulio (2018) provided evidence that the reinforcing value of a token increases with generality. Using human participants, they manipulated generality by making tokens exchangeable for either salty snacks only, several types of food and drink options, a gift card. They found that as generality increases, the relative reinforcing value of different types of tokens also increases.

In a similar study with pigeon subjects, DeFulio et al. (2014) assessed the reinforcing value of three different types of tokens under conditions of water deprivation: food tokens (exchangeable for only food), water tokens (exchangeable for only water), and generalized tokens (exchangeable for food or water). In this study, subjects produced more generalized than specific tokens across several increasing token production schedules, demonstrating a higher reinforcing efficacy for generalized tokens. Overall, increasing generality increases reinforcer value.

In addition to manipulating the number of types of back-up reinforcers available in a token menu, the exchange value of items on the menu (i.e. price) can be changed in several ways. The first way to manipulate price is to change the token exchange schedule. This could be done by changing the overall price of every item on the menu or by manipulating the price of specific menu items. With certain menus all item may be the same price. For example, one token may be exchangeable for one bag of chips or 15 minutes of iPAD time. The price of specific menu items can also be manipulated by having highly preferred items that are more expensive (outings, money) and easily accessible items with low cost (pack of gum, small bag of candy) on

the same menu. For both manipulations the cost of the back-up reinforcers may be relative to the actual cost of the items. For example, a bag of beef jerky or trail mix would likely cost more than chips. The individual implementing the token economy may then have to determine the relative cost of one type of reinforcer to another (e.g. the value of 15 minutes of iPAD time compared to a bag of chips). Increasing the token exchange schedule typically decreases response rate. Malagodi et al. (1985) increased the token exchange schedule by increasing the number of tokens a rat had to deposit into the terminal in order to access food. Response rate decreased substantially when the token exchange schedule was increased from an FR-1 to and FR-4.

A second way to manipulate price is as a product of exchange production schedule. As discussed previously, this can be done by using a token board. While the token exchange schedule may be an FR-1 ("I have earned all of my tokens"), increasing the exchange production requirement serves a similar function in that it increases the number of tokens required to access the back-up reinforcer. For example, by filling a token board with five spaces that produces 15 minutes of iPAD time, the cost of the iPAD time is five tokens, without any manipulation of the token exchange schedule. Overall, increasing price should decrease value of tokens as it will increase the response effort in one of the component schedules. Responding may become resistant to changes in price depending on the items included in the menu.

Exchange schedules can also be manipulated by adding delays to terminal reinforcement. For example, Leon et al. (2016) compared three conditions of delayed reinforcement: delayed access to food reinforcement, delayed token delivery, and immediate token delivery with a delayed exchange period. Response frequency decreased in the first two conditioned when food and tokens were delayed. Responding persisted in the conditioned where token delivery was immediate, but exchange periods were delayed. This study models clinical applications of token

economies where tokens are delivered immediately contingent on a required response, but participants may have a limited number of exchange periods per day, or exchange on specific days of the week.

Measuring Token Utility

To build the most robust token economy possible, a clinician must create conditions under which tokens have high reinforcing value, and exchanging is frequent. Producing these ideal conditions requires knowledge of how the availability of back-up reinforcers and token component schedules individually and conjointly affect token value. While the token component schedules and menu manipulations described above may have profound effects on the reinforcing value of tokens, developing a robust method for measuring reinforcer value is important for building an effective token economy.

The most common assessments of reinforcer value are preference assessments. While there are several types of preference assessments, trial-based assessments are very prevalent in practice. In trial-based assessments the participant is presented with one, two, or several stimuli concurrently across several trials. Participants respond to the presented stimuli and the responses are recorded and a hierarchy is typically formed (DeLeon & Iwata, 1996). There are many trial-based preference assessments, including paired stimulus and multiple stimulus presentations. In a paired stimulus preference assessment, a participant is presented with two concurrently available stimuli and selects the most preferred. The process is repeated until all possible pairings have taken place and a preference rank is formed (Fisher et al., 1992). While the paired stimulus assessment produces reliable outcomes, it is often very time consuming depending on the number of items being assessed. This problem has been approached by including several stimuli in an array, rather than only two. In a multiple stimulus preference assessment, the entire array of

stimuli are presented simultaneously with the participant selecting the most preferred stimuli. This procedure can be modified where the selected stimuli remains in the array and the stimuli not selected stimuli are replaced with a new set (multiple stimulus with replacement [MSWO]) or where the selected stimulus is removed from the array and the remaining stimuli are selected to produce a hierarchy (Windsor, Piche, & Locke, 1994). Preference assessments are useful in that they are reliable tools that produce results comparable to how a participant would respond in a real scenario where the high preferred stimulus is serving as a reward. However, it is important to note that preference assessments and reinforcer assessments are entirely different procedures. The former attempts to predict how a participant would respond if that stimulus were presented contingent on a target behavior, the latter has the actual contingency in place.

The next procedure that has frequently been used to assess reinforcer value is the progressive ratio (PR) procedure. A significant different between a PR task and preference assessment is that a PR task includes an actual measure of reinforcer value. Preference assessments can produce results that are indicators of value. For example, a student may select chips over candy in a paired stimulus preference assessment. Therefore, we can assume that chips will work as more powerful reinforcers than candy. PR tasks, however, include an actual contingency to assess value. In a PR task, participants are presented multiple fixed ratio schedule requirements where the number of responses required to earn a reinforcer is steadily increased upon each consecutive session (Hodos, 1961). The point at which the participant ceases to respond on the schedule is called the breakpoint. In terms of reinforcer value, the higher the breakpoint for a specific reinforcer, the larger its reinforcing value. In terms of token economies, breakpoints of tokens with different back-up reinforcers associated with them can be compared. For example, Tan & Hackenberg (2015) assessed PR breakpoints for three different types of

tokens: ones that could be exchanged for food only, water only, and both food and water (generalized). In their study, tokens exchanged for food had the highest breakpoints, followed by water and generalized tokens. Traxler & DeFulio (In Prep) measured the effects of token generality on reinforcing value using a progressive ratio task. Rather than the dichotomous approach of Tan and Hackenberg (food or water vs. generalized tokens) Traxler & DeFulio (In Prep) measured generality using a graded approach by increasing token generality across three different conditions. In a progressive ratio task, the most generalized tokens produced the highest breakpoints, providing evidence that generality increases the value of conditioned reinforcers.

Behavioral economics may provide a more thorough assessment of reinforcer value via the demand analysis. The purpose of a demand analysis is to provide a measure of the reinforcer value. These measures would allow a practitioner to predict the reinforcing efficacy of a stimulus (Hursh & Silberberg, 2008). This is done by plotting how the consumption of a commodity changes as the effort to acquire it increases. In typical laboratory studies this would be primarily a plot of the number of reinforcers earned in each session as the response requirement increases. A token economy study may plot the number of tokens earned as a function of price (e.g. Traxler & DeFulio, In Prep). When consumption is plotted as a function of price those plots are called demand curves (Hursh & Silberberg, 2008). Below is an exponential model of demand (Hursh, 2014) with consumption being plotted as a function of an FR schedule requirement.

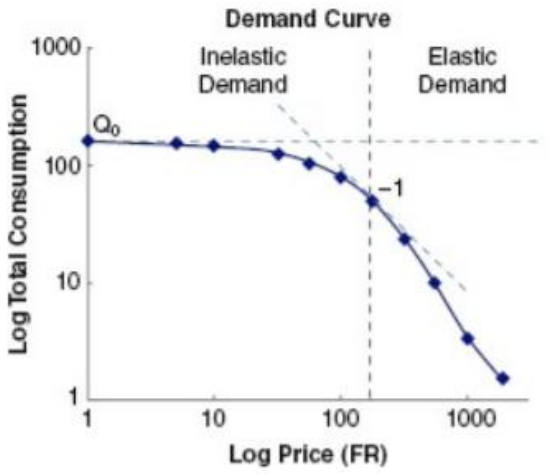


Figure 1. Reinforcer consumption curve (retrieved from Hursh, 2014)

There are two major components of the exponential model of demand used to create demand curves: the 1) consumption when the commodity is free (Q_0) or the point in which demand for an item would be the highest; plus 2) the elasticity of demand. Elasticity of demand refers to the rate of decrease of consumption with increases in price, also called sensitivity to price (Hursh & Roma, 2016). Consumption is typically considered to be elastic when a one percent increase in price results in greater than a one percent decrease in consumption. In other words, the consumption of a commodity is very sensitive to price increases. Consumption of luxury commodities may be an example, in which increases in price may result in people looking for less expensive alternatives. Conversely, consumption may be inelastic when a one percent increase in price results in less than a one percent decrease in consumption. In this case, consumption is not very sensitive to changes in prices. This typically occurs with essential items such as gasoline, though the decrease in price of electric vehicles may increase elasticity. The two components discussed form the basics of the exponential model of demand:

$$\log Q = \log Q_0 + k(e^{\alpha P} - 1). \quad (1)$$

In this equation Q references the demand for a reinforcer and Q_0 being consumption when price is zero. K refers to a constant that specifies the logarithmic range of data. Alpha (α) equals rate of change in elasticity, and P is the price determined by either a dollar amount or a schedule requirement.

Figure 2 contains the amount of response output required to support the demand curve displayed in figure 1. This figure displays two other important variables: P_{MAX}, which is the price in which maximum responding occurs (Hursh and Winger, 1995); and O_{MAX} which is the maximum amount of responding that occurs at P_{MAX} (Roma et al., 2015)

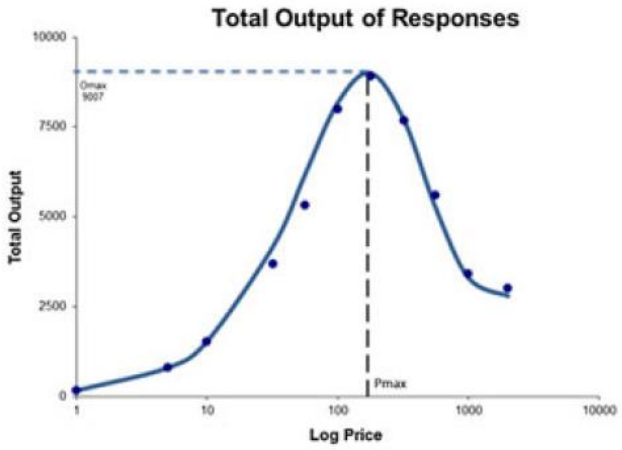


Figure 2. Response output to meet levels of demand shown in figure 1 (Retrieved from Hursh, 2014)

The demand curves have led to the creation of a measure of reinforcer strength. This is termed the essential value of a reinforcer, which is the reinforcing value of a commodity independent of reinforcer size, calculated as the inverse of alpha (α; rate of change in elasticity). Essential value provides an opportunity to compare reinforcers and is a product of the equation below:

$$EV = 1/ (100 \bullet \alpha \bullet k^{1.5}) \quad (2)$$

The Hypothetical Purchase Task

While demand analyses are robust assessments of reinforcer value, they are often intensive and require many trials to test the demand for commodities at various prices. The number of sessions can grow exponentially as multiple independent variables are introduced. To alleviate this issue, researchers often use a self-report measure of consumption via the Hypothetical Purchase Task (HPT) to measure demand (Roma, Hursh, & Hudja, 2015). The HTP was originally designed to measure hypothetical consumption of commodities at various prices and has roots in substance abuse to measure consumption of certain drugs (heroin, cigarettes) at several price points. However, since its inception the HPT been extended to measure consumption in several different contexts including sexual health (Strickland et al., 2020), the consumption of alcohol (for a meta-analysis see Kiselica, Webber, & Bornovalova, 2015), the use of steroids (Pop et al., 2010), food (Epstein et al., 2010), the internet (Broadbent & Dakki, 2015, fuel (Reed et al., 2014), and gambling (Weinstock, Mulhauser, Oremus, & D'Agostino, 2016) to name a few. The HTP is primarily used to model real world consumer behavior when obtaining true data on consumption is unethical, not practical, or otherwise impossible (Jacobs & Bickel, 1999).

While the "hypothetical" component of an HPT may be concerning due to predictive validity issues, there has been extensive research demonstrating accurate comparisons to actual experiences. Amlung et al. (2012) compared responding during a hypothetical alcohol purchase task to a task with actual alcohol rewards and found a close correspondence between the two for both demand for and consumption of alcohol (Wilson et al., 2016). Other studies comparing self-reported consumption of various drugs to hypothetical purchase tasks have found strong

reliability and validity, including alcohol consumption (Murphy and MacKillop, 2006; Murphy et al., 2009), cocaine (Bruner and Johnson, 2014), and cigarettes (MacKillop et al., 2008).

Using an HPT for coffee consumption as an example, the general format of an HPT is as follows: 1) the participant is provided an image of a cup of coffee; 2) a description of the hypothetical scenario is provided along with a timeframe (e.g. "Imagine that you are thirsty for a cup of coffee. The following scenario asks how many coffees at various price points you would buy in one month"); and 3) participants are informed of any assumptions they must make, such as a lack of coffee available outside of those prices, only you can consume the coffee, you cannot save or sell the coffee, and any information about their income. This is an example of a task in which the participant must calculate the number of commodities they would purchase at various prices. This is popular task involving everyday use items in which people may buy multiple (Roma, 2015). However other tasks involve calculating the proportion of participants that would make a purchase at various prices. In this situation, instead of participants providing a number, they instead are given a sliding bar to indicate the probability of them making a purchase at the selected price-points. This is popular for single use items or items not purchased in large quantities (Roma, 2015). Both formats allow for the systematic manipulation of several independent variables by changing the hypothetical scenario in addition to the price points of each commodity. This provides an opportunity to assess how multiple variables may affect demand separately, and to test independent variable interactions.

There are several variables that affect the demand for a commodity. Many of these variables involve the availability of alternative sources of reinforcement (Hursh & Roma, 2016). A method to quickly affect demand is to control the availability of specific reinforcers outside of the experimental sessions. Participants are said to be operating under an open economy when

they have free access to reinforcers outside of experimental sessions. When the access to a reinforcer is restricted to only the session, they are operating under a closed economy (Hush, 1980; Hursh & Roma, 2016). These types of economies can create drastically different response patterns as a product of change in price and response requirements (Hush & Roma, 2016). In a closed economy, especially with primary reinforcers, demand much less elastic than in an open economy (Hursh, 1978). Token economy research in laboratory settings typically involves a mix of both closed (Andrade & Hackenberg, 2017) and open (Yankelevitz et al., 2008) economies, where participants are maintained at 80% of free feeding weight outside of experimental sessions (open) or have their access to back-up reinforcers limited to experimental sessions only (closed). There is little clinical research evaluating the effects of outside session reinforcer availability on responding in a token economy. While it could be assumed that responding would follow a similar pattern as with primary reinforcers, this hasn't been demonstrated empirically. There may be several additional variables, such as generality, that would affect elasticity of demand in a closed vs open economy.

While token systems typically follow the same general format there are many individual and potentially interacting variables that affect their utility. When designing token economies, it is reasonable to suspect that the above variables that affect demand for general commodities have the same effect on demand for tokens and the back-up reinforcers associated with them. Additionally, token economies have several components that may also have differential effects on demand, including the token production, exchange production, and token exchange schedules. Each of these schedules increase's the response effort in some way, from increasing the work requirement to earn a token, to exchange it, and the cost of the back-up reinforcer at the moment of the exchange. It is also important to note that tokens acquire their value through their

association with the back-up reinforcers. Therefore, those items should be carefully considered as they likely moderate the relationship between the three component schedules and demand.

Purpose

The purpose of this study was to measure the effects of token menu manipulations and token component schedules on demand for tokens.

This was the first time a hypothetical purchase task had been used to measure demand for tokens. Prior to completing the study, two pilot studies were conducted to assure that the main study will be designed in a way to promote accurate responding on the hypothetical purchase task. The first pilot tested the presentation order of three variations of the hypothetical purchase task, each being worded in a slightly different way. This was done to assess for sequencing effects that may confound the results of future presentations of a hypothetical purchase task. The second pilot was used to assess which of three possible variations of the hypothetical purchase task results in demand that is most sensitive to changes in price, which was used for the main study.

In the main study three independent variables were manipulated to assess their effects on demand for tokens. The first manipulation was the token production schedule. To be applied to a hypothetical purchase task, participants must respond to an array of commodity prices. To model a typical token economy the token production schedule was selected to be included in the price array. Nine token production schedule values were selected to provide the minimum number of responses to produce a price sensitive measure of demand (Roma, 2015). The second independent variable was the types of items on the menu. Four menu categories were selected based on common preference and functional assessment tools and to assess the demand for each type of back-up reinforcer. The “mixed” category was included as a measure of generality. The

final independent variable in the main study was the number of items on the menu. Finally, the possible number of menu items included were three, six, and twelve. There are two purposes for including these values. First, it served as a model for typical menu sizes in applied settings and encompasses potential extreme values. Second, it allows for the manipulation of menu categories, primarily an even mix of all three categories. The manipulation of type and number of back-up reinforcers available provides robust information on generality, a principal determinant of reinforcer value Token exchange schedules were considered. However, it would not be possible to include nine token exchange schedules on the HPT with a menu larger than one item.

General Methods

The follow section provides an overview of study methods that were uniform across all three experiments. Once covered, each experiment will have its own procedure, results, and discussion.

Recruitment

Amazon Mechanical Turk (MTurk) was used for subject recruitment. MTurk is an online labor market where "requesters" (researchers or organizations requesting information) can post jobs for "workers" (the employees). In this experiment, the work we requested is called a Human intelligence task ("HITs"), which can vary from answering surveys, to more involved tasks such as transcribing audio and conducting web searches. In this case participants answered a survey. The study survey was a HIT on the MTurk website was accessible by eligible participants.

All participants were recruited from January 15th, 2021 until March 1st, 2021. Participants were recruited rapidly, with about 50 participants completing the survey within about two hours

of the survey being uploaded onto MTURK. 100 Participants completed Pilot 1, and 50 participants completed Pilot 2. For the main study, the survey was distributed until 50 acceptable data sets were acquired, which required 82 respondents.

To be eligible to participate in the study participants must have 1) been located in the United States; 2) had a Human Intelligence Task (HIT) approval rate of at least 95%; and 3) had at least 100 approved HITs (Robinson et al., 2019). Increasing the number of prior approved HITs decreases the potential participant pool. This increases the probability that participants will have prior experience with Hypothetical Purchase Tasks which may affect their responding. Additionally, given that a large percentage of HITs are completed by a small percentage of workers, there is a chance that most studies that use MTURK are using the same participant pool when the prior approved HIT requirement is at or over 500. Having the 95% approval rating helped improve the likelihood that participants were filling out their responses truthfully and not randomly (Peer et al., 2014).

Informed Consent

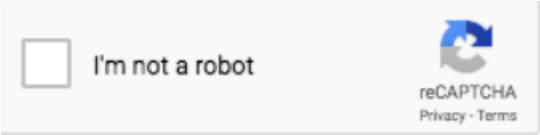
The informed consent process was completed on the first page of the survey. The last part of the informed consent form stated "By accepting and submitting this HIT, you agree to these terms. Clicking "Next", participants were agreeing to the terms of the informed consent. No signature was required.

Participant Screening

Attention Checks. Three times during the survey participants were asked a question unrelated to study procedures with the purpose of decreasing the likelihood of including participant's responses in the data analysis that were random or the product of an automated

program (also called “bots”). These questions had clear correct answers, examples including a simple delay discounting task (“Would you rather have $1,000 now or $1 in one year?”) or simple trivia (“Which city is not located in the United States” with Beijing being the correct answer). Participants that get any of these questions incorrect were compensated, but their responses were not included in the final data analysis.

Use of Captcha Verification Question. The final question of the survey that participants were required to complete prior to getting a completion code was a Captcha verification question (Completely Automated Public Turing Test to tell Computers and Humans Apart).



The completion of the Captcha verification question took approximately one second to complete and was designed to prevent automated programs from completing the survey and accessing study payment. Only participants that complete the Captcha question were compensated.

Training Questions

To gain access to the study survey, participants were required to complete a brief training to assure that they properly understood how to use the basic components of the Qualtrics survey. Participants were provided the marketplace menu that they would encounter during the main portion of the study survey. Above the menu was the following script:

“*Imagine you earn money by building widgets. You can use your money to buy the things you see on the menu below. You can see 12 items on the menu. Each item has a cost to the right of the item. For example, 30 minutes of gym time costs $8. A glass of beer or wine costs $4. Let's make sure you understand the basics of how the menu works.”*

Following the script and the presentation of the menu, participants were given three multiple-choice comprehension questions to assure they understood the basics of the menu. For example, participants were asked "*If you had to create 1 widget to earn a dollar, how many widgets would you need to make to purchase a single serving of chips, hot chocolate, or candy?"* (correct answer is 2 widgets). Using survey validation, if participants selected an incorrect answer, they were told they were incorrect and to select the correct answer. Participants could not move on to the next section of the survey until they answered all three questions correct.

Comprehension Test

Upon completing the training questions participants were provided the following script:

> "*Now it is time for the comprehension test. Please remember that if you do not pass the comprehension test, you will be excluded from the study and ineligible for compensation*."

Upon clicking "next" participants were presented with a four-question comprehension test. To pass the comprehension test, participants must have answered 100% of the questions correct. Participants that met this criterion continued to the survey, while participants that did not were told "You did not pass the screener question. When you press next, you will be exited from the survey." Below is an example of a screening problem:

How many dollars (type the number only) would you need to earn to purchase hot coffee??

General Procedure

After completing the comprehension check participants were instructed to complete a specific number of hypothetical purchase tasks, depending on the study, which followed the general format described below.

Hypothetical Purchase Task (HPT) Questionnaires

The hypothetical purchase task (HPT) followed the same general procedure of previous research using HPTs, primarily the methods described by Roma et al. (2015). However, these HPTs included a menu of options rather than a single commodity. The following was included in each question on the HPT:

1. Written description of the hypothetical purchase scenario including

    a) The cost of each item on the menu

    b) A general description of the type of menu item

2. Assumptions and limitations (availability of these items outside of the questionnaire, participants cannot sell anything after they purchase it, items are for their consumption only).

3. Nine fixed ratio schedule requirements with room for participants to indicate the amount of work they would complete when given a specific response requirement to earn payment.

Data Analysis

While data analysis was conducted in a manner consistent with prior demand research using hypothetical purchase tasks, the interpretation of those outputs has unique differences. This is primarily because the dependent measures have been modified. First, participants were asked to state how much work they would complete given a specific response requirement. Typical demand research using an HPT asks participants to state how many of a commodity they would purchase at various prices. Therefore, all y-axes are stated in terms of consumption. In the present study, however, all axes are described in terms of production of widgets, tokens, or dollars. Demand intensity (Q0) and elasticity (alpha) are also interpreted differently. In this

study, Q0 refers to the production of tokens, widget, or dollars when the response requirement to do so is zero. Alpha refers to the rate of change of elasticity of production.

The differences in the interpretation of the basic components of the demand model results in changes in the interpretation of essential value, which is typically used to compare the strength of reinforcers. In this study, with all other variables being held constant, essential value can be used to compare the strength of each menu.

All data that was gathered through MTurk was automatically saved in the Qualtrics database and was exported as a Microsoft Excel file. Graphpad Prism was used for all exponential demand analyses and the generation of the exponential model. All participant data was pooled for the purposes of the demand analysis.

Demographic Questionnaire

Upon the completion of the main study questionnaire, participants answered a brief demographic questionnaire. The questionnaire included the following questions, with participants selecting the response that most applies to them:

1) What is your gender?

2) What age group are you in?

3) What is your ethnicity?

4) What is the highest education you have completed (four categories)

5) What is your primary profession

6) What is your household income range (six categories)

Design

A within subject, experimental survey was used to measure the effects of various independent variables on the hypothetical consumption of back-up reinforcers on the menu.

Primary Outcome Variable

The primary outcome variable that used in this study was the number of widgets a participant was willing to make when presented with a response requirement to earn one dollar.

Data Inclusion/Exclusion Criteria

A participant's responses were included in the primary data analysis if 1) they passed the four-question comprehension test; 2) 100% of the survey was completed as displayed on Qualtrics's Data Analysis; 3) they submitted a survey code on Amazon MTURK.

A participant's responses were excluded if 1) they generated highly stereotyped responses (e.g. answered 12 for every response), which is a clear indication of completing the survey as fast as possible; 2) they provided physically impossible value amounts (e.g. over 1440 widgets made, *Pilots 1 & 2 only*); 3) they were matching the number of widgets required to earn a token (e.g. if the token production schedule was FR1, they put 1, and did the same for all other schedules; 4) they failed the screener but somehow made it through the survey (Pilot 1 only); or 5) they failed any of the three attention checks. Below are examples of responses that meet the exclusionary criteria discussed above:

	Token Production Schedule								
Exclusion Criteria	1	5	10	25	50	100	200	400	800
1	1	1	1	1	1	1	1	1	1
2	1440	7200	14400	36000	72000	144000	248000	416000	512000

3	1	5	10	25	50	100	200	400	800

Pilot 1

The purpose of Pilot 1 was to assess for sequencing effects when manipulating the order participants complete the hypothetical purchase task (HPTs). The results of this pilot would inform the order in which participants would be presented each of the three HPTs in Pilot 2.

Procedure

We manipulated the sequence participants completed two HPTs, with two groups receiving the HPTs in different order. To assure that reliable demand curves could be created using an HPT in this study, a 12-item menu with each type of reinforcer category was used. The prices of each back-up reinforcer were relative to real world cost and were held constant for the entire study. If a sequence effect was observed (e.g. demand curves for the same task depended on the presentation order) the hypothetical purchase tasks were presented in a randomized order for Pilot 2. Below is a table describing each possible hypothetical purchase task:

Name	How was the HPT Framed?	Example Question on HPT
HPT1	Tokens	How many **tokens** would you earn in a day if you had to create 1 widget to earn a token?
HPT2	Dollars	How many **dollars** would you earn in a day if you had to create 1 widget to earn a dollar?
HPT3	Widgets	How many **widgets** would you make in a day if you had to create 1 widget to earn a dollar?

Participants that completed HPT1 were first presented a menu and hypothetical purchase task that was framed using standard tokens. Participants were told "*Imagine you earn tokens by building widgets. You can use your tokens to buy the things you see on the menu below*." All items on the menu were presented in terms of the number of tokens or dollars required to purchase each back-up reinforcer. Each question on the HPT was written in the format above.

Participants that completed HPT2 were given a menu with identical items. However, rather than using the word "tokens", the scenario was framed using dollars. Participants were told "*Imagine you earn money by building widgets. You can use the money to buy the things you see on the menu below*." All items on the menu were presented in terms of the price (in dollars) required to purchase each item. All questions on the HPT were framed using dollars, rather than tokens (e.g. "*How many dollars would you earn in a day if you had to create 1 widget to earn a dollar*?")

Participants that completed HPT3 were first presented a menu identical to HPT2. The framing of the scenario was also identical to HPT2 (all items on the menu were presented in terms of the number of tokens required to purchase each back-up reinforcer). However, rather than answering how many dollars they would make, participants were asked how many widgets they would make given a wage. For example, participants on the FR-5 portion of the HPT were asked "*How many widgets would you make in a day if you had to create 5 widgets to earn a dollar?"*

Participants in group 1 received HPT1, then HPT 2. Participants in group 2 received HPT 2, then HPT 3. Clear sequencing effects were observed when analyzing the demand for money when HPT 2 was presented first, rather than second (see Results). Therefore, the order of HPT presentation was randomized, and no further piloting of sequence effects was required.

Duration

During pilot 1, participants in each group were required to complete the training questions, comprehension exam, two hypothetical purchase tasks, and the demographics survey. The average time to complete Pilot 1 for group 1 was 11.5 minutes (*SD* = 5.25 minutes) with a range of 3.5 to 22 minutes. For group two the average time to complete Pilot 1 was 10.5 minutes (*SD* = 5.15 minutes) with a range of 5 to 26.5 minutes.

Results

In Pilot 1 the exponential demand model provided a moderate fit to demand curves for group 1, resulting in R^2 values of .72 for the Token framed task, and .72 for the Money framed task. For group 1, the exponential model of demand yielded a weak fit for both the Money and Widget framed tasks (R^2 =.46, .40). While the demand model for all three framed HPTs provided a moderate fit at best, demand was highest in the second presented condition for both groups (see table 2). For group 1, demand intensity (Q0) was highest and alpha values were lowest in the money framed HPT. The opposite was true for group 2, where demand intensity was lowest in the money framed HPT and alpha was highest.

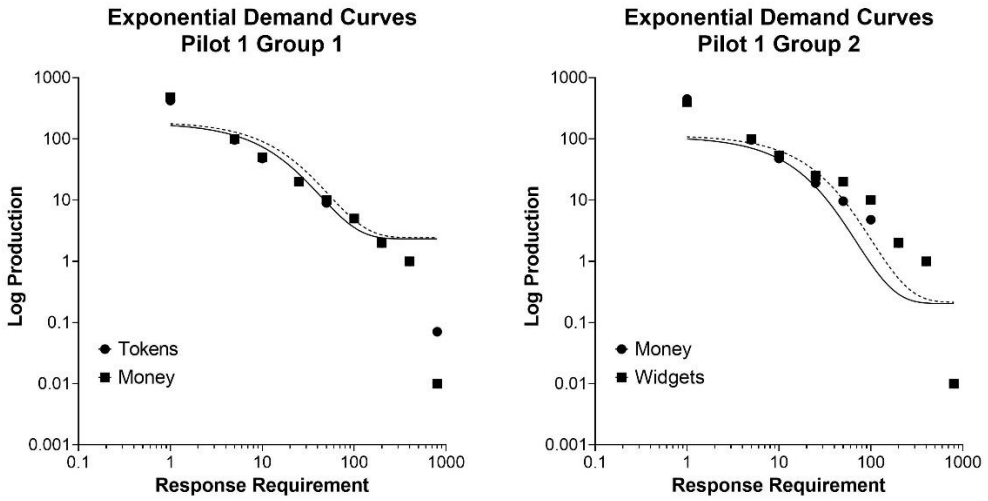


Figure 3. Results of demand analyses for groups 1 and 2 for Pilot 1. Participants in group 1 completed a hypothetical purchase task that was framed using tokens, followed by an HPT that was framed using money. Participants in group 2 first completed an HPT that was framed using money, followed by a similar task that asked participants to estimate how many widgets they would make.

Table 1. Best fit values for groups 1 and 2 for Pilot 1

	Group 1		
Condition	**Q0**	**k**	**alpha**
Tokens	183	1.9	0.00013
Dollars	192	1.9	0.000098
	Group 2		
Condition	**Q0**	**k**	**alpha**
Dollars	109	2.7	0.00013
Widgets	115	2.7	0.000085

Discussion

Overall, the alpha increased for the "how many dollars would you earn" condition when it was presented first, which shows less resistance to changes in the production schedule. However, it was unclear whether this was due to a sequencing effect, or a product of demand being lower for the second group overall. Given that randomization would assure that sequencing effects didn't confound the results of Pilot 2 and the main study it was decided to randomize the presentation of study HPTs for Pilot 2 and the main study. Additionally, practically speaking, sequencing effects won't be a factor when designing a token economy, unless it is continuously

being changed. Participants would rarely be presented a sequence of several different token menus in an applied setting.

Pilot 2

Procedure

The purpose of Pilot 2 was to compare demand for each of the three HPTs described above when presented in a randomized order, with the goal to select one of the HPTs for use in the main study; and troubleshoot any issues discovered during the analytics of Pilot 1 data. One group was used in this study, with each participant receiving the three HPTs in a random order. See the above Table 1, or the Pilot 1 section above, for a summary of each HPT. To determine the HPT to use in the main study, the HPT whose framing produced the highest essential value was selected.

Duration

During pilot 2, participants were required to complete the training questions, comprehension exam, three hypothetical purchase tasks, and the demographics survey. The average time to complete Pilot 2 was 11.5 minutes (*SD* = 5.25 minutes) with a range of five to 19.5 minutes.

Results

In Pilot 2 the exponential demand model provided a weak fit to demand curves across all conditions, yielding an R^2 value of .49, .57, and .48 for the token, dollars, and widgets conditions, respectively. The widgets condition produced the highest demand intensity (Q0 = 133) and an alpha of .00018. The Tokens condition produced the lowest demand intensity (Q0 = 87) and the highest demand elasticity (alpha = .0003).

Table 2. Demand values for Pilot 2

	Condition		
	Tokens	**Dollars**	**Widgets**
Q0	87	102	133
k	3.2	3.2	3.2
alpha	0.0003	0.00022	0.00018

The demand curves for each condition across all participants are included below. On average, dependent variables decreased steadily as the response requirement to earn money or tokens was increased. This trend continued until the final four production schedules, where responding decreased to near zero across all conditions. A slightly less elastic demand curve can be observed when hypothetical purchase tasks were framed using the production of widgets, rather than the earning of dollars or tokens.

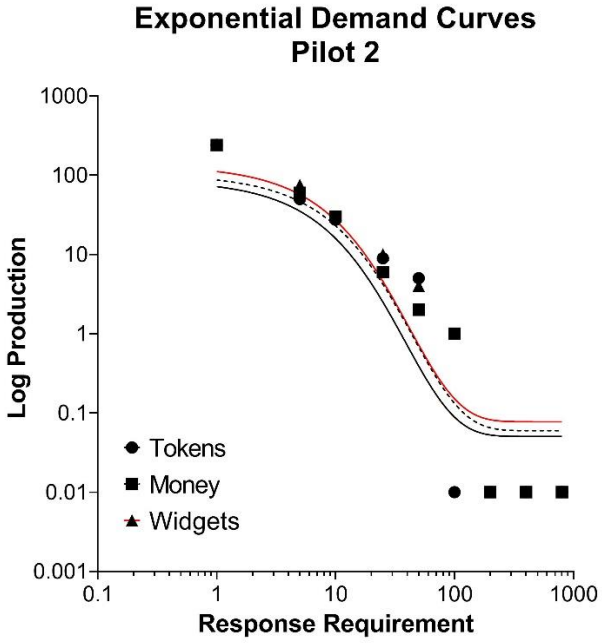


Figure 4. Exponential demand curves for Pilot 2. Participants received each HPT in a random order.

The primary analysis used to determine the HPT presentation for the main study was essential value (EV). Below are the EVs for all three pilot conditions. HPTs framed in terms of how many widgets a participant would make produced an essential value of 9.71, which was the

highest among all conditions and used for the main study. The next highest EV was for the HPTs framed using dollars (EV = 7.94) followed by the tokens frame (EV = 5.82). Overall, the value of the menu was highest in the widget condition.

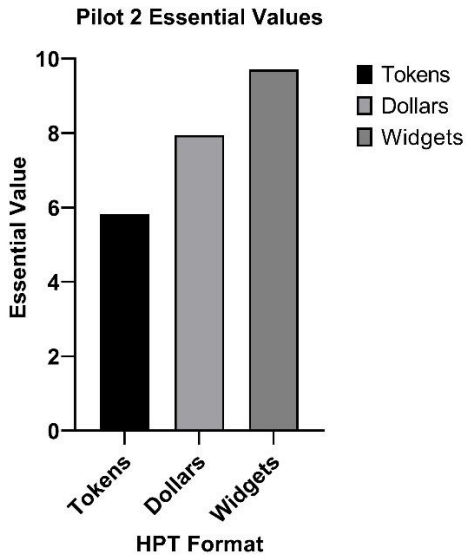


Figure 5. Essential values across all conditions for Pilot 2

Discussion

In Pilot 2 the HPT that was framed "how many widgets would you make" had the highest Q_0, value. However, the HPT framed "How many tokens would you earn" had the lowest demand elasticity. Instead of using one of those values to dictate the task that was used in the main study, essential value was used as our determinant, as it is a function of Q_0, Alpha, and k values. As shown in figure 3, the HPT framed "how many widgets would you make" had the largest essential value. Based on this result all HPTs used in the main study were framed this way.

Main Study

Procedure

Participants completed hypothetical purchase tasks as described in in the "General Methods" section. All HPTs were framed asking participants "how many widgets would you make" when given a response requirement to earn a dollar. However, the main study included the addition of the two primary independent variables, which were menu manipulations.

Menu Manipulations

Each description of the HPT scenario included a marketplace the participant had access to, which contained a manipulatable variety of commodities at fixed prices. The participant indicated how many widgets they were willing to make at various token production schedules.

1) Type of items on the menu- Menu items were in three main categories: 1) Food; 2) Activities; and 3) Tangibles. Rather than being provided with specific items, participants were given a broader description of the menu item to account for their preferences. For example, the three-item food menu included 1) A single serving of chips; 2) An 8-inch sub, burger, or similarly prepared food; 3) a bottled water. Menu types were either only food, only activities, only tangible items, or an even mix of all three.

2) Number of items on the menu- The size of the menu participants will select from will was manipulated by containing three items, six items, or twelve items.

Token Production Schedule

The main body of the hypothetical purchase task used nine token production schedules as a measure of "price." This was a statement of the number of responses a participant is required to

make to earn a dollar and was based on the results of Pilot 2. Participants indicated the number of widgets they were willing to make each token production schedule.

Below is the general format for each HPT:

Imagine you have volunteered your day to participate in a project in which you are building widgets. You earn money that can only be used at the projects market. You can use your money to buy the things you see on the menu below:

Item	Cost
A single serving of chips	$2
A bottled water	$2
8-inch sub, burger, or similarly prepared food	$8

Assumptions:

1. You have other NO ACCESS to any of these items other than the ones available above
2. You cannot sell or give away any of these items
3. It takes about 1 minute to make a widget

When answering the questions below, **consider the menu of items available and think about how hard you would want to work. You can stop working at any time and go home**. There are no "right" or "wrong" responses. Please answer all questions honestly, thoughtfully, and to the best of your understanding, as if you were actually in this situation.

How many widgets would you make in a day if you had to create **1 widget** to earn a dollar?

There were 108 combinations of all independent variables in this study including the nine TP schedules in each HPT. In addition to the nine TP manipulations, there are multiple levels of

two independent variables. A breakdown of the IV manipulations are included in the table below:

IV	Level								
1) Token Production Schedule (Fixed Ratio)	1	2	3	4	5	6	7	8	9
2) Type of items on the menu	All Food		All Activities		All Tangibles		Even Mix		
4) Number of items on the menu	3			6			12		

Each participant underwent every possible independent variable combination. There was one HPT for each combination of type of items on the menu, and number of items on the menu. Therefore, each participant completed 12 HPTs. Each HPT had nine questions, each question corresponding to one of the nine token production schedules, totaling 108 questions to complete all study procedures.

Duration

During the main study participants were required to complete the training questions, comprehension exam, 12 hypothetical purchase tasks, and the demographics survey. The average time to complete the main study was 38 minutes (*SD* = 35 minutes) with a range of 13 to 107 minutes.

Data Analysis

Additional exclusion criteria primarily those described by Stein et al. (2015) were applied to the data set prior to conducting the demand analysis. These included excluding data based on

nonsystematic trend, bounce, and reversal. The following formula was applied to each completed hypothetical purchase task:

$$\text{Delta Q} = (\log Q_1 - \log Q_n) / (\log P_n - \log P_1)$$

Delta Q is the change in the quantity of widgets a participant would make. Q_1 and Q_n number of widgets made at the first and last price, respectively. Finally, P_n and P_1 and are the last and first token production schedules. Stein et al. (2015) suggested a criterion of X = .025 to compare to Delta Q. If delta Q was less than X, the data was considered nonsystematic as it detected less than a .025 log reduction in consumption for each log unit increase in token production schedule. Below are examples of systematic and nonsystematic data sets.

	Fixed Ratio Schedule									
	1	**5**	**10**	**25**	**50**	**100**	**200**	**400**	**800**	Delta Q
Responses	100	60	40	35	25	15	8	4	2	0.586
	500	250	250	600	300	600	200	400	800	-0.07

Following the Stein et al. (2015) method for removing nonsystematic data two demand functions were applied to the data set. Koffarnus et al. (2015) developed an exponentiated model of demand. This model was created to address the issue of inputting zero consumption values in the demand model. Zeroes cannot be included on a logarithmic scale. Log 0 is undefined and is not a real number, so cannot be included in the exponential demand model. Koffarnus et al. (2015) highlights several approaches to the problem of zero consumption values. They include omitting the zeroes from the analysis, replacing the zeroes with smaller values, like .01, and only analyzing larger group models, rather than individual demand curves.

When applying the exponential model to a dataset, zero values are automatically emitted. This means that the dependent measure for each participant top when their consumption levels reach zero. This results in a demand curve like below:

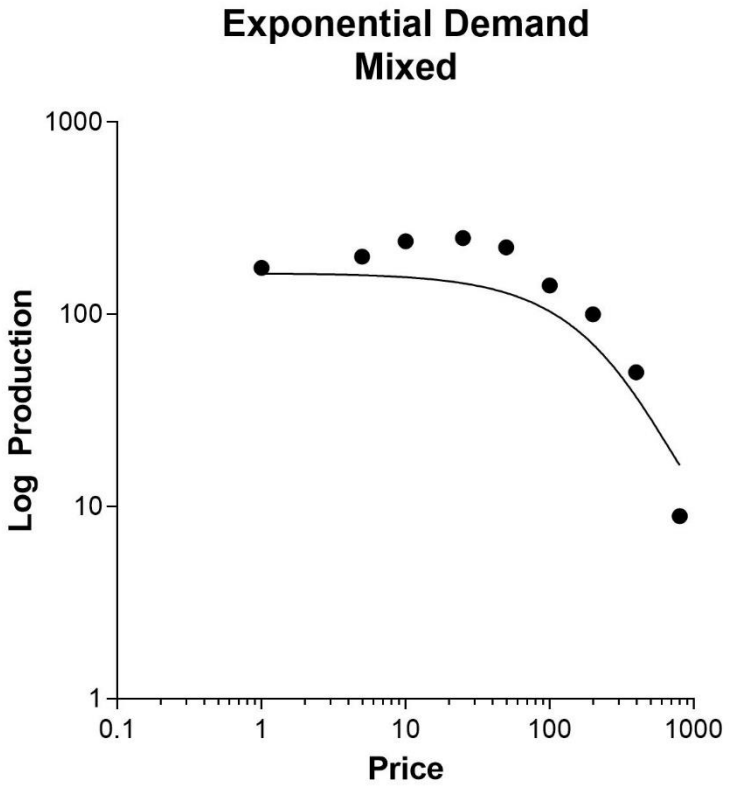


Figure 6. Exponential demand example

In this scenario, elasticity of demand is deflated due to leaving out lower production values. Additionally, R2 values decrease. In the example above $R^2 = .13$. Additionally, leaving out zeroes is excluding important participant data (Koffarnus et al., 2015).

One approach to this issue is replacing all zero values with a very small number, typically 0.01. This prevents zero values from being replaced. Below is an example of a data set that is

modified by replacing zeroes with small values, which was used in the exponential demand model for this study:

Type	1	5	10	25	50	100	200	400	800
With Zeros	975	870	300	250	200	0	0	0	0
Zeros replaced	975	870	300	250	200	0.01	0.01	0.01	0.01

The model proposed by Yu et al. (2014) and later by Koffarnus et al. (2015) includes a modified equation that is an exponentiated version of the previous demand equation that can include zero values into the analysis. The exponentiated model is expressed below where both the demand intensity and elasticity sides of the equation are raised to the power of 10:

$$Q = Q_0 * 10^{k(e^{-\alpha Q_0 C} - 1)} \quad (3)$$

The exponentiated model was also included in the analysis of the present study. Models were compared based on resulting Q^0, alpha, k, and R^2 values.

To calculate essential value, participants' responses remaining after exclusionary criteria was applied were transferred to Graphpad Prism to undergo the exponential demand analysis. All Q_0, k, and alpha values resulting from the analyses were transferred to a free Microsoft Excel calculator developed by Kapland & Reed (2014) where essential value was calculated.

All statistical analyses were completed using Graphpad Prism. A one-way ANOVA was conducted to assess for differences between essential values across all 12 IV combinations. In this analysis, each combination was treated as its own column with the means of each column being compared in the ANOVA. Prior to conducting the ANOVA, outliers were excluded using

the ROUT method, which is an outlier identifier available on Graphpad Prism. Contingent on the discovery of statistically significant differences between the mean essential values for each condition, a Tukey multiple comparisons test was used to assess individual differences in means between each condition.

A two-way ANOVA was conducted to assess for main effects on essential value across reinforcer type and number of items on the menu. Additionally, interaction effects were calculated between both independent variables using the two-way ANOVA.

Results

It was predicted that increasing the number of items on the menu would slightly increase demand when reinforcer type is held constant. Therefore, participants would make more widgets on the 12-item menu than the three-item menu. This would be indicated by a higher Q_0 and a low alpha. It was also predicted that the increase in demand as a product of menu size would be constant across each of the three reinforcer types. Demand curves were expected to be similar for each of the three back-up reinforcer categories. Finally, it was predicted that the combination that would result in the highest demand will be the mixed, 12 item condition. This would be the largest menu, with several reinforcer categories. The combination with the lowest demand was predicted to be the 3 item, primary reinforcer category.

Figure 4 contains the exponential demand curves for the aggregate median widgets produced during all study conditions. Overall, demand increased as the menu size increased. Demand intensity was lowest in the three foods condition ($Q_0 = 78$) followed by the six foods condition (Q0 = 105). Intensity was highest in the six tangibles condition, followed by the 12

tangibles condition. Demand elasticity decreased as the menu size increased for all four menu types.

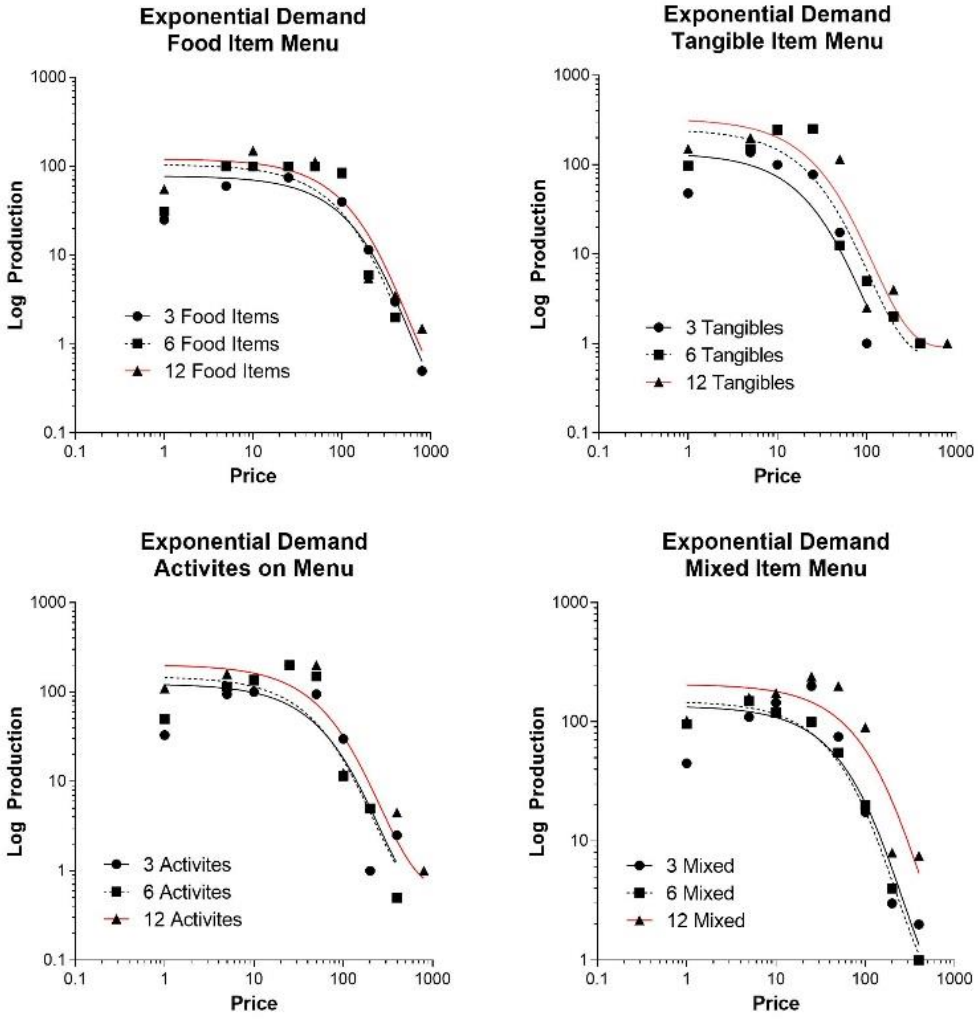


Figure 7. Exponential demand curves for aggregate median data.

Figure 5 contains demand curves for pooled data. Like the median data displayed in figure 4 above, demand increased as the menu size increased. However, all 12 demand models produced a low R^2 value, indicating the model produced a weak fit. Demand intensity, elasticity, and essential value increased as the menu size increased for all 12 menu types. Alpha and Q_0 was

highest and lowest respectively, in the three-primary reinforcer condition. Overall, Q_0 was lowest in the three primary reinforcer conditions. In all conditions the production of widgets decreased to near zero levels at the 100-widget response requirement.

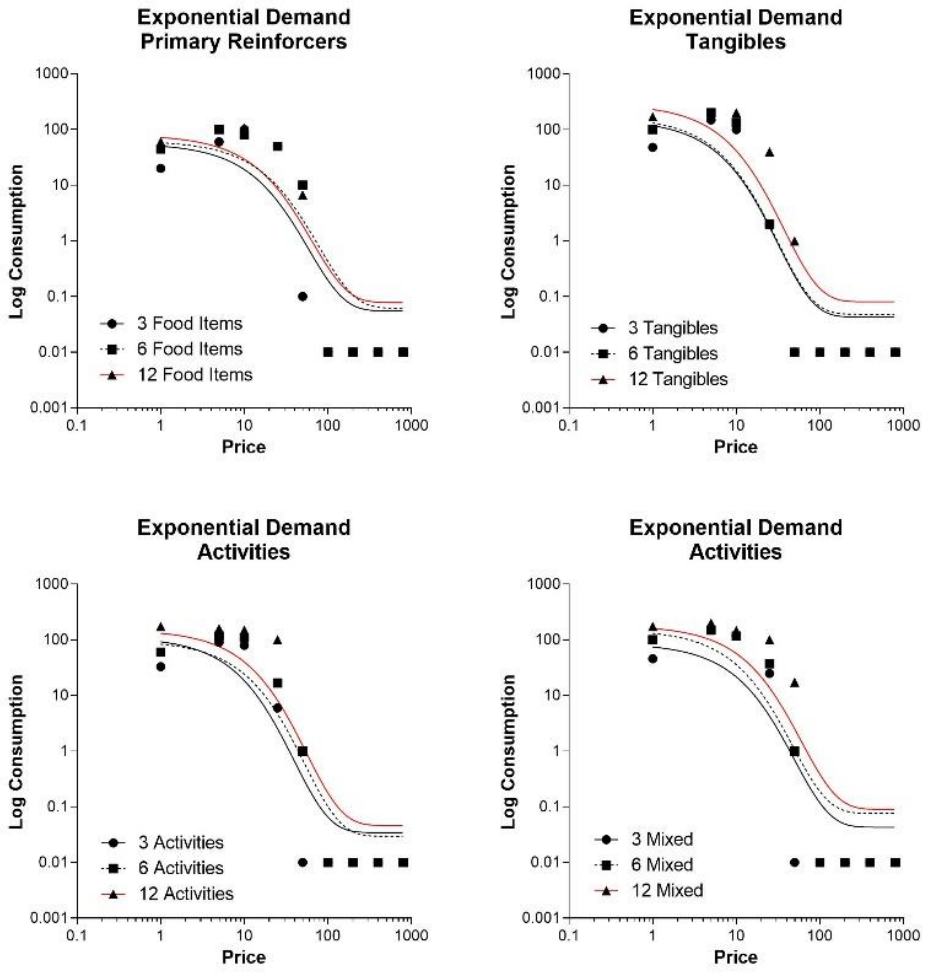


Figure 8. Demand curves for pooled participant data after removing trend violations using the Stein et al. (2015) method.

Figure 6 includes the exponentiated demand curves for pooled participant data in the main study using the methods employed by Koffarnus et al. (2015). Overall, the exponentiated model provided a good fit for the dataset, with a median R^2 value of .92. Demand for all conditions was highest in the 12-item menu categories. However, these results should be

interpreted with caution. As shown in table x, there was a high amount of variation between k and alpha values. Additionally, the exponentiated model of demand was not able to fit a curve for the 6 tangible and 6 activity conditions (R^2 = -.20; 0, respectively). Visually, the 6 activity and 6 tangible production values appear to follow a similar curve to the other 6 item conditions. However, the exponential model did not provide a fit.

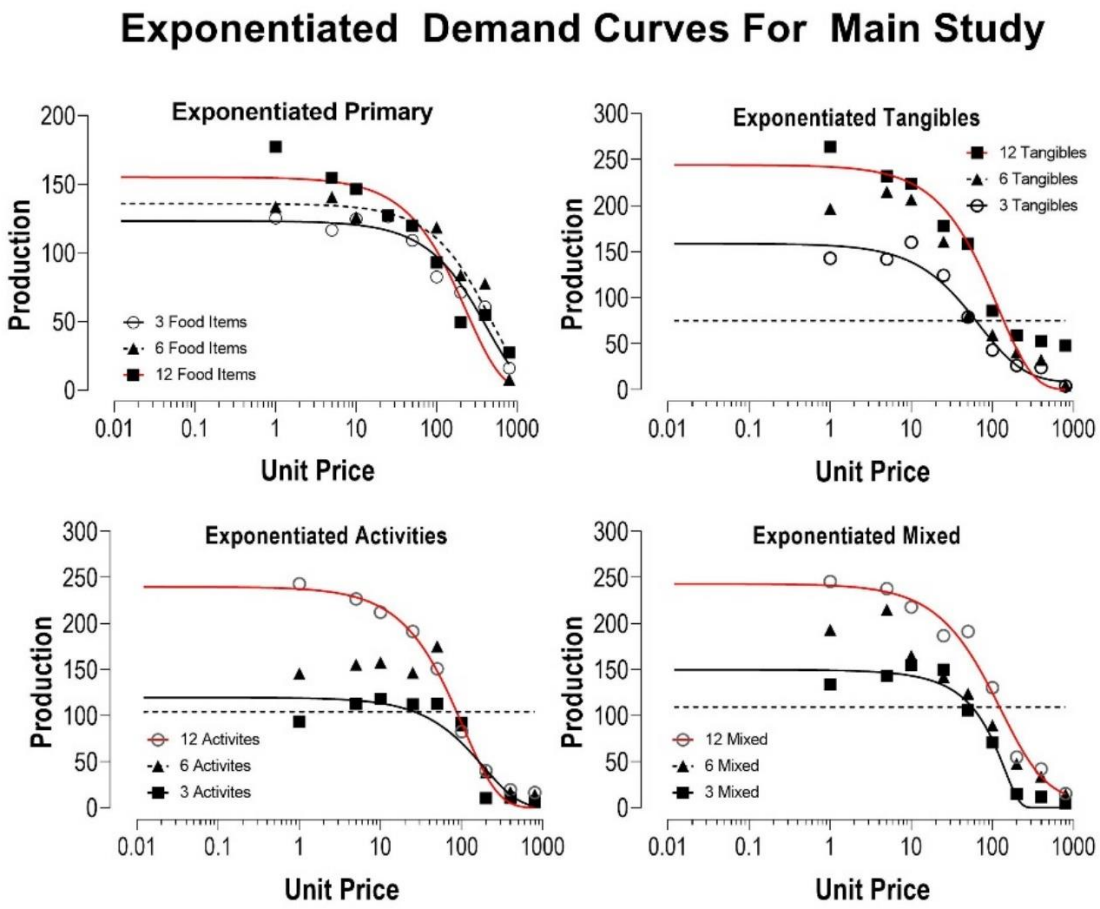


Figure 9. Exponentiated demand curves for main study (Koffarnus et al., 2015)

Table 3. Demand variables exponentiated demand analysis

Values	**3 Primary**	**6 Primary**	**12 Primary**
Q0	123.2	135.8	155.1
k	~ -751.2	~ 1908	~ -3958
Alpha	~ -1.107e-008	~ 3.466e-009	~ -2.997e-009
R squared	0.9498	0.9389	0.8971
Values	**3 Tangibles**	**6 Tangibles**	**12 Tangibles**
Q_0	158.7	75.04	244.1
k	1.281	~ -0.0001318	~ -3351
Alpha	2.945e-005	~ 8.317e-010	~ -4.539e-009
R squared	0.9635	-0.2048	0.9015
Values	**3 Activities**	**6 Activities**	**12 Activities**
Q_0	119.7	~ 73.01	239.8
k	~ 1869	~ -0.1547	~ -2213
Alpha	~ 1.029e-008	~ 6.171	~ -7.848e-009
R squared	0.8657	0.000	0.9902
Values	**3 Mixed**	**6 Mixed**	**12 Mixed**
Q_0	149.8	109.4	243.1
k	-0.1929	~ 6.781e-007	1.461
Alpha	-6.364e-005	~ -3.531e-012	9.079e-006
R squared	0.9718	-0.004082	0.9812

Essential value calculations for the main study are presented with both the inclusion and exclusion of major outliers. Figure 6 contains the average essential values for each of the 12-menu combination, with outliers included. On average, the essential value increased as the menu size increased. A one-way repeated measures ANOVA was conducted to compare the effects of menu size and type on essential value. There was not a significance difference detected between conditions at the $p>.05$ level for all conditions [$F(11,549) = .5299$, $p = .8835$]. As displayed on figure 6 and table 3, there was immense amount of variation in the data set. This was especially apparent in the 3-primary item ($M = 1189$; $SD = 6475$) and 12 tangible item menus ($M = 635.4$; $SD = 3007$). There were also a vast range in essential values when outliers were included, from 1591 in the 6 tangibles condition, to 42670 in the 3-primary condition.

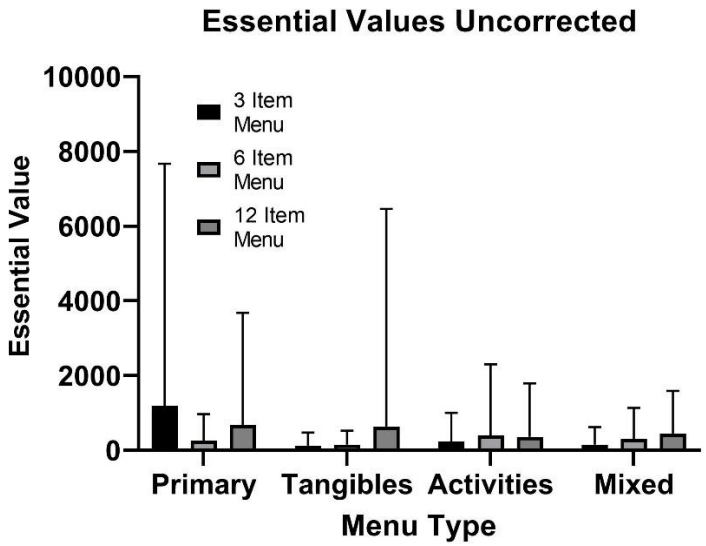


Figure 10. Essential values of uncorrected data for each category.

Table 4. ANOVA summary table raw essential values

RAW EV ANOVA Table	SS	DF	MS	F	P value
Treatment	45701490	11	4154681	F (11, 549) = 0.5299	P=0.8835
Residual	4.3E+09	549	7840208		
Total	4.35E+09	560			

Table 5. Raw essential values descriptive statistics

Category	Min	Max	Range	Mean	Std. Deviation
3 Primary	0.1487	42670	42670	1189	6475
6 Primary	0.1585	4134	4134	252.1	712.6
12 Primary	-2067	18645	20712	667.5	3007
3 Tangible	0.1487	2144	2144	119.4	353.6
6 Tangible	0.1506	1591	1591	149	369.9
12 Tangible	-12232	37253	49486	635.4	5821
3 Activity	-148.7	3775	3923	235.3	762.6
6 Activity	0.1537	12629	12629	390.8	1912
12 Activity	-1021	8951	9973	357.9	1430
3 Mixed	0.1509	2717	2717	149.4	471.1
6 Mixed	0.1608	3775	3775	304.1	826.8
12 Mixed	0.1576	6304	6304	446.5	1139

Figure 7 contains average essential values for all study conditions after outliers were removed using the ROUT method. Of the 561 total essential values collected in the main study,

104 (18.5%) were identified as outliers and removed from the data set. Essential value increased as the menu size increased for all conditions. A one-way repeated measures ANOVA was conducted to compare the effects of menu size and type on essential value. A significant difference was detected between conditions at the p>.05 level for all conditions [F(11,445) = .6.880, p < .001]. In the primary reinforcers only condition, essential value was the highest in the 12 item condition (M = 24.85) and lowest in the three item condition (M = 17.70). This trend was similar for all study conditions. When sorted by essential value (table 11) the 12 mixed condition had the highest essential value (61.01), followed by 12 tangible (49.75) condition. The lowest essentual values were observed in the 3 primary (14.72) and 3 tangible conditions (17.70). A similar trend was observed for all other primary determinants of demand, including Q_0 and alpha values (see tables x, and x).

A sharp decrease in essential value variation was observed after removing major outliers. According to table 6, the highest standard deviation observed was in the 12 mixed condition (*M* = 61.01; *SD* = 60.45) and the lowest in the 3 tangible condition (*M* = 14.72; *SD* = 13.4). The range of means also decreased significantly when removing outliers, with the largest range of

data occuring in the 12 mixed condition (Range = 215.5) and lowest in the 3 tangible condition (Range = 45.95).

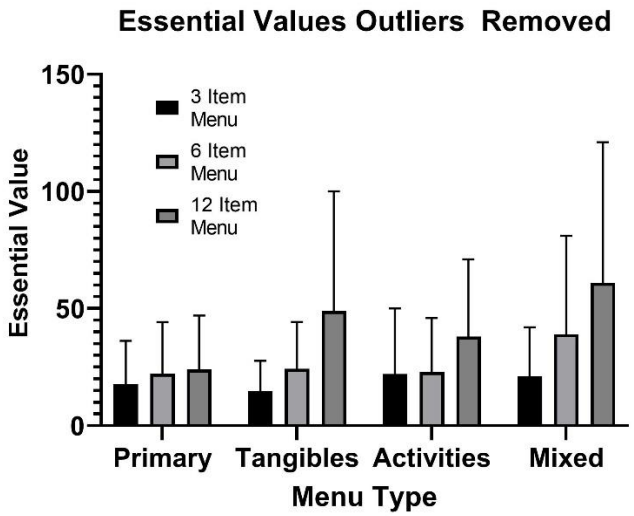


Figure 11. Essential values of pooled data after the removal of outliers using the ROUT method.

Table 6. ANOVA summary table for corrected essential values

	SS	DF	MS	F	P value
Treatment	86498	11	7863	F (11, 445) = 6.880	P<0.0001
Residual	508632	445	1143		
Total	595130	456			

Table 7. Corrected essential value descriptive statistics

Category	Min	Max	Range	Mean	Std. Deviation
3 Primary	0.1487	70.1	69.95	17.7	18.51
6 Primary	0.1585	103.4	103.2	22.22	22.67
12 Primary	0.164	86.45	86.28	24.85	23.86
3 Tangible	0.1487	46.1	45.95	14.72	13.4
6 Tangible	0.1506	85.18	85.03	24.27	20.83
12 Tangible	0.1576	182.1	182	49.75	51.87
3 Activity	0.1487	98.14	97.99	22.59	28.78
6 Activity	0.1537	104.8	104.6	23.08	23.76
12 Activity	0.155	144.7	144.6	38.82	33.24
3 Mixed	0.1509	86.45	86.3	21.29	21.39
6 Mixed	0.1608	192.9	192.8	39.61	42.03
12 Mixed	0.1576	215.7	215.5	61.01	60.45

A Tukey multiple comparisons test was used to assess for significant differences between each condition, with the statistically significant results of the test displayed on table 7. As predicted the largest mean differences were observed between the 12 mixed category and three item menus of the other categories. The largest mean difference was observed between the 12 mixed and 3 tangibles condition (46.29; $p < .0001$) and between the 12 mixed and three primary conditions (43.32; $p < .0001$). Overall, there appeared to be a lack of significant mean differences between the three menu item categories, all which produced non-statistically significant different essential values.

Table 8. Significant results from the Tukey multiple comparisons test

Tukey's multiple comparisons test	**Mean Diff.**	**95.00% CI of diff.**	**P Value**
3 Primary vs. 12 Tangibles	-32.05	-57.01 to -7.085	0.0018
3 Primary vs. 12 Mixed	-43.32	-68.99 to -17.64	<0.0001
6 Primary vs. 12 Tangibles	-27.52	-52.49 to -2.561	0.0168
6 Primary vs. 12 Mixed	-38.79	-64.47 to -13.12	<0.0001
12 Primary vs. 12 Mixed	-36.16	-62.23 to -10.10	0.0004
3 Tangibles vs. 12 Tangibles	-35.02	-59.80 to -10.25	0.0003
3 Tangibles vs. 12 Mixed	-46.29	-71.78 to -20.80	<0.0001
6 Tangibles vs. 12 Tangibles	-25.48	-50.08 to -0.8807	0.0347
6 Tangibles vs. 12 Mixed	-36.75	-62.07 to -11.43	0.0002
12 Tangibles vs. 3 Activities	27.16	2.727 to 51.59	0.0152
12 Tangibles vs. 6 Activities	26.67	1.702 to 51.63	0.0246
12 Tangibles vs. 3 Mixed	28.45	3.674 to 53.23	0.0099
3 Activities vs. 12 Mixed	-38.42	-63.58 to -13.27	<0.0001

Table 8 continued.

6 Activities vs. 12 Mixed	-37.93	-63.61 to -12.26	0.0001
3 Mixed vs. 12 Mixed	-39.72	-65.21 to -14.23	<0.0001

Table 8 displays the results of the two way analysis of variance which was conducted to compare the main effects of menu size and categories and the interaction effect between both independent variables on essential value. A significant difference was detected between conditions at the p>.05 level for both the number of menu items [$F(2,445) = 20.71$, $p < .001$] and reinforcer category [$F(3,445) = 6.206$, $p = .0004$]. Additionally a statistically significant interaction between reinforcer category and number of menu items was observed [$F(6,445) = 2.139$, $p = .0479$].

Table 9. Results of two-way ANOVA of essential values

ANOVA Table	**SS**	**DF**	**MS**	**F**	**P value**
Interaction	14670	6	2445	F (6, 445) = 2.139	P=0.0479
Number of Menu Items	47335	2	23668	F (2, 445) = 20.71	P<0.0001
Reinforcer Category	21282	3	7094	F (3, 445) = 6.206	P=0.0004
Residual	508632	445	1143		

Table 10. Demand Variables Pooled Data

Values	**3 Primary**	**6 Primary**	**12 Primary**
Q0	56	63	80
k	3	3	3
Alpha	0.0003	0.0002	0.00019
R squared	0.34	0.33	0.37
Mean Essential Value	17.7	22.22	24.85
Values	**3 Tangibles**	**6 Tangibles**	**12 Tangibles**
Q_0	152	171	286
k	3.5	3.5	3.5
Alpha	0.00021	0.00019	9.4e-005
R squared	0.43	0.48	0.38
Essential Value	14.72	24.27	49.75

Table 10 continued.

Values	**3 Activities**	**6 Activities**	**12 Activities**
Q_0	113	97	151
k	3.5	3.5	3.5
Alpha	0.00022	0.00019	0.00012
R squared	0.41	0.51	0.47
Essential Value	22.59	23.08	38.82
Values	**3 Mixed**	**6 Mixed**	**12 Mixed**
Q_0	87	154	181
k	3.3	3.3	3.3
Alpha	0.00023	0.00014	9.1e-005
R squared	0.45	0.39	0.37
Essential Value	21.29	39.61	61.01

Table 11. Results sorted by Q0

Category	**Q_0**
3 Primary	56
6 Primary	63
12 Primary	80
3 Mixed	87
6 Activity	97
3 Activity	113
12 Activities	151
3 Tangibles	152
6 Mixed	154
6 Tangibles	171
12 Mixed	181
12 Tangibles	286

Pooled data results sorted by Q_0 (smallest to largest). Larger values indicate higher demand intensity.

Table 12. Results sorted by alpha

Category	**Alpha**
12 Tangibles	9.4e-005
12 Mixed	9.1e-005
12 Activities	0.00012
6 Mixed	0.00014
12 Primary	0.00019
6 Activities	0.00019
6 Tangibles	0.00019
6 Primary	0.00020
3 Tangibles	0.00021

Table 12 continued.

3 Activities	0.00022
3 Mixed	0.00023
3 Primary	0.00030

Results sorted by demand elasticity (smallest to largest). Smaller values indicate more inelastic demand.

Table 13. Study conditions sorted by essential value

Category	**Essential Value**
12 Mixed	61.01
12 Tangibles	49.75
6 Mixed	39.61
12 Activities	38.82
12 Primary	24.85
6 Tangibles	24.27
6 Activities	22.59
3 Activities	22.59
6 Primary	22.22
3 Mixed	21.29
3 Primary	17.70
3 Tangibles	14.72

Study conditions sorted by essential value. Higher EVs indicate higher reinforcing efficacy of the back-up reinforcers on the menu.

Table 14. Study conditions sorted by Pmax

Category	**Pmax**
6 Primary	15.31
3 Primary	14.43
12 Mixed	13.23
12 Primary	13.21
6 Activities	11.54
12 Activities	9.69
3 Mixed	8.77
3 Activities	8.45
6 Mixed	8.04
12 Tangibles	5.51
3 Tangibles	5.43
6 Tangibles	5.31

Study conditions sorted by Pmax which is the price in which maximum responding occurs (Hursh and Winger, 1995)

Table 15. Study conditions sorted by Omax

Category	Pmax
12 Mixed	696.40
12 Tangibles	646.29
6 Mixed	428.55
12 Activities	414.04
12 Primary	399.45
6 Primary	315.35
6 Activities	299.03
3 Tangibles	252.05
6 Tangibles	252.05
3 Activities	244.66
3 Primary	239.67
3 Mixed	232.13

Study conditions sorted by Omax which is the maximum amount of responding that occurs at PMAX

Table 16. Study conditions sorted by R Squared

Category	R-Squared
6 Activities	0.51
6 Tangible	0.48
12 Activities	0.47
3 Mixed	0.45
3 Tangibles	0.43
6 Mixed	0.39
12 Tangibles	0.38
12 Primary	0.37
12 Mixed	0.37
3 Primary	0.34
6 Primary	0.33
3 Activities	0.41

Larger R-squared values indicate a higher goodness of fit for the exponential model when applied to the final dataset and suggest that the menu explains a larger amount of the variation of the dependent variable.

Qualitative Responses

Over 200 qualitative responses were gathered across both pilots and the main study. While these responses were not formally classified, participants' comments generally fell under the following categories:

1. General comments of study approval (e.g. *"good", "Fair", "Nice", "Good Study*"). These types of comments were common. This is likely because participants on average have a higher likelihood of being paid when they leave comments in open ended sections of surveys.

2. Statements about their willingness to work as the response requirement increased (e.g. *"I was thinking about the amount of work and how hard it would be. I wanted to put myself in a scenario and asked myself how much I was willing to do" OR "As the amount of work to make a dollar increased, my determination to work decreased.").*

3. Statements involving math. This was typically either how they used math to solve the problems (*I just tried to do the math inside my head and I answered accordingly")* or how they were frustrated with having to do math (e.g. *"I was thinking oh great math.")*

4. Philosophical statements about the task ("*Do I really need the items presented. Can I live without the presented items? Would the work for the items be worth it*?")

5. Very detailed analyses of their own responding ("*For the first round for a new table I found the $60/hr to be generous so I thought about what I would want from the table. If it included durable goods, then I figured I would just put in an 8-hour day for those wages. As the hourly rate declined, I thought more about what I would really want. A subscription service kind of won out here at lower wages if it was available. I wasn't willing to work for anything once the wages got below $6 an hour.)*

Demographics

Table 12 contains important demographic information for participants included in the main study. The mean age for participants in the main study was 41.1 (SD = 11.1) with a mean household

income of $44,335.09 (SD = $39,219.58). Most participants were Caucasian and held at least a bachelor's degree.

Table 17. Demographic information for main study

Variable	Category	Number	Percent
Gender	Male	32	65.31%
	Female	15	30.61%
	No-binary	2	4.08%
Age	18-24	2	5.71%
	25-34	15	42.86%
	35-44	12	34.29%
	45-54	14	40.00%
	55-64	5	14.29%
	>65	1	2.86%
Ethnicity	Caucasian	35	71.43%
	African American	7	14.29%
	Asian	5	10.20%
	Hispanic/Latin American	1	2.04%
	Prefer Not to Answer	1	2.04%
Highest Education Completed	Less than High School degree	0	0.00%
	High School or equivalent	8	16.33%
	Some college but no degree	6	12.24%
	Associate degree	6	12.24%
	Bachelor's degree	21	42.86%
	Master's degree	7	14.29%
	Doctoral Degree	0	0.00%
	Professional Degree (JD, MD)	1	2.04%
Household Income	<$25,000	15	30.61%
	>25,000 to < $50,000	16	32.65%
	>$50,000 to < $75,000	9	18.37%
	>$75,000 to < $100,000	4	8.16%
	>$100,000 to < $125,000	2	4.08%
	>$125,000 to < $150,000	0	0.00%
	>$150,000	1	2.04%
	No Response	2	4.08%

Discussion

In this study, essential value increased as the number of items on the menu increased for all four study conditions. A statistically significant interaction was observed between the number of menu items and reinforcer category, on essential value. Additionally, the mixed 12-item menu produced the highest essential values, and the 3 primary reinforcers menu produced the lowest essential values. The three lowest essential values were observed in the primary reinforcer menu conditions. However, the tangible item condition produced significantly higher essential values than the primary and activity reinforcer categories, which was unexpected.

The results of this study align with prior studies involving generalized conditioned reinforcement. This includes recent research using human participants (e.g., Traxler & DeFulio, In Prep) demonstrating an increased reinforcer value with increases in generality; and non-human research like DeFulio et al. (2014) where subjects produced more generalized tokens rather than ones exchangeable for specific types of primary reinforcement. Additionally, the present study extends the findings of previous research by compartmentalizing generality. To reiterate, according to Skinner (1953) behavior maintained by a generalized reinforcer is likely to be under the control of multiple states of deprivation. Simply increasing the size of the menu only partially accounts for more generalized tokens. The interaction effect observed in the present study provides evidence that multiple variables play a role in menu efficacy. To provide the most generalized menu possible, multiple different types of back-up reinforcers should be presented in the menu, rather than simply increasing the number available.

The present study was the first assessment of token efficacy using a hypothetical purchase task. Increasing token production schedule requirements resulted in changes in responding like FR schedule manipulations in progressive ratio tasks (i.e. Hodos, 1961; Traxler

& DeFulio, In Prep) and previous demand studies using token economies (i.e. Tan & Hackenberg, 2015). This was demonstrated in Figure 4, in which demand curves were generated using median data without the application of the Stein et al. (2015) method. Token production schedule requirements decreased responding when the data was pooled, but nonsystematic data (e.g. results in which demand increased as token production schedule requirements were increased) were removed. This resulted in filtering data that did not conform to the general pattern of an inverse relation between the number of widgets made response requirement to earn one dollar. In addition to consistency with previous research, results of this study are also consistent with the law of demand which states that consumption decreases as price of a reinforcer increases (Stigler, 1954).

The current study was also the first time the hypothetical purchase task was modified to assess response output, rather than frequency of purchase. The literature on assessing hypothetical output is limited. The existing literature has been an extension of delay discounting (Madden & Bickel, 2010) and most commonly employs effort discounting tasks. Effort discounting models the decrease in value of a commodity as the effort to acquire it increases (Mitchell, 2004). In an effort discounting procedure, participants are required to make choices between a small outcome that is available with little response requirement, and a large outcome that is only available after completing a more effortful task (Malesza et al., 2019). While current behavior analytic methods may better measure reinforcer value (i.e. progressive ratio schedules), effort discounting tasks are the only other assessment of hypothetical response output. Malesza et al. (2019) found that the results of hypothetical effort discounting tasks mirrors those of real effort tasks. This is consistent with literature using hypothetical purchase tasks compared with actual purchases (e.g. Amlung et al., 2012).

Implications for Behavior Therapy

The results of this study have several implications when it comes to the treatment of target behavior for individuals receiving behavior therapy. When designing a token economy, back-up reinforcers that are related to multiple motivational operations should be included whenever possible. This is especially true with primary reinforcers, that will control responding under a limited number of motivating operations, even when the menu size is increased. This recommendation is based on the interaction effect that was observed between certain types of reinforcers, such as activities and tangible items, that appear to control behavior under more MOs. This is likely because activities (e.g. time on social media) can result in contact with other social reinforcers that aren't available when consuming only primary reinforcers, like food and water. Other tangible reinforcers (e.g. a tank of gas, a piece of art, clothing items) can also result in access to other types of reinforcement.

Despite the evidence provided in this study that primary reinforcers have limited efficacy compared to tangible reinforcer based and mixed menus, other reinforcer categories can be modified for individuals with limited preferences. Restricted preferences are more common for individuals living with Autism Spectrum Disorder and is included in the diagnostic criteria (APA, 2000). Activity based reinforcers can be altered to involved functionally equivalent activities related to the problem behaviors being targeted. For example, in Kahng et al. (2003) a token economy in which tokens were exchangeable for food removal was employed for an individual who engaged in food refusal. This could also be applied to other function-based activities including the removal of demands (for escape and avoidance-maintained problem behavior) in the form of homework passes or time out of the classroom (Gillis & Pence, 2015). However, there is little research comparing generalized and function-based back-up reinforcers

on the menu. Future research in this area would contribute to our understanding of token menus and potentially fading out token economies more effectively.

An additional consideration that should be made is the size of the menu. As displayed on figure 7, the type of back-up reinforcers on the menu has a limited effect on demand when the menu is small. In the present study, this was likely partially explained by a lack of preference assessments being completed prior to the presentation of the menu. For example, if the only tangible item on the three item menu was a new video game or a tank of gas, and the participant didn't own a car or play video games, only two items on the menu may have any value. While preference assessments are a common component of the functional behavior assessment process, their necessity increases when using smaller token menus.

The results of this study also provide important information when it comes to effectively fading token economy interventions. Fading is a critical step in the systematic removal of a token economy intervention. The propensity to relapse is common in any behavioral intervention, whatever the target, when the intervention involves the use of extrinsic reinforcers. From a theoretical perspective, this happens because the function of the target behavior is to acquire a token. For example, in a basic token economy intervention, a child may earn tokens for completing math problems. When the ability to earn tokens is withdrawn, completing math problems no longer produces a token, and baseline patterns of responding re-emerge. This is not a criticism of token economies. Assuming the effects would persist if the token economy was removed would not be aligned with our basic understanding of stimulus control. There is no reason to think a behavior would persist if it was no longer producing the reinforcing stimulus. However, there is evidence that gradually fading token contingencies can promote maintenance. For example, Philips et al. (1971) was able to maintain desirable target behaviors at Achievement

Place when token delivery was decreased to 8% of its original rate. There are several ways in which a token economy could be faded. They involve the gradual increase of the token production or exchange production schedules. More research is required to understand the relationship between exchange production increases and demand. However, it is typically recommended to increase the exchange production schedule when fading a token economy (Hackenberg, 2018). Doing so is taking advantage of a token's ability to bridge the gap between a target behavior and terminal reinforcement. By increasing the exchange production schedule, the token has already been earned, a delay is being added to the exchange period, which may not decrease demand as substantially as token production increases. Some practitioners may increase the token production schedule gradually when fading a token economy. When doing so, it is important to have a larger menu with several back-up reinforcer categories. This would create less elastic demand for tokens that would be less sensitive to the changes in token production schedules. If only three primary reinforcers are included on the menu, increases in token production schedules may have a profound effect on demand, and decrease the therapeutic effect of the token economy.

Utilizing Amazon Mechanical Turk

The study was the first to assess token demand using a hypothetical purchase task. It was also the first study to incorporate hypothetical effort, rather than consumption, using a hypothetical purchase task. There were several observed benefits to distributing the survey online via MTurk. Primarily, running this study in person where participants have experience with each token production schedule would be extremely cumbersome. To obtain a similar dataset, participants would have to undergo 108, 1-hour sessions, one for each level of each

hypothetical purchase task. This would not be feasible in an in-person setting. Using MTurk, however, participants were able to input 108 production values in about 51 minutes, on average and were paid at a rate slightly over $10 per hour. A similar pay rate would have cost approximately $1000 per participant if done in person. Amazon also provides access to a large participant pool of about 250,000 workers worldwide who have completed at least one HIT (Robinson et a., 2019).

While there are many benefits to using Amazon Mechanical Turk, they come with several caveats. The first major consideration when using MTurk the determination of the 95% approval rating and a minimum of 100 HITs completed (Peer et al., 2014), which was employed in this study. According to Robinson et al., about 35% of MTurk workers have completed fewer than 100 HITs. Additionally, participants with fewer than 1000 previous HITs completed make up a small fraction of participants in MTurk research. This means that researchers using MTurk are sampling from a small portion of MTurk employees and often surveying the same group of people. Therefore, there are limitations to the ability to generalize results to the general population. Additionally, there is limited evidence that increasing worker qualifications results in more reliable data (Robinson et al., 2019). Initially, the previous HIT criteria were 500, but was decreased to 100 to gather more experimentally naive participants.

Even with carefully selected inclusion criteria participants may be behaving in a way to maximize earnings, rather than always providing reliable answers. Several MTurk web extensions, such as “MTurk Suite, and “Stax” exist to help a worker maximize the amount they can earn per hour. When conducting within-subject analyses, the impact of each participant’s responses has a more significant effect on descriptive statistics. Therefore, careful attention is required to potential outliers that could create noise in the data set. Even though there are several

considerations that need to be made when using MTurk, most Workers provide truthful answers and rationale. About 70% of participant data was usable for purposes of the demand analysis.

Controlling the Economy using Rules

The descriptions of the hypothetical purchase tasks evolved from the first pilot to the main study. This was primarily because participants were treating the initial hypothetical purchase tasks as math problems with correct answers, rather than their willingness to work for the items on the menu. This was apparent in qualitative responses. For example, several participants in the first pilot made statements similar to "I subtracted the number of minutes it took to make the widgets from 1440 (number of minutes in 24hrs) and multiplied that by the number of widgets made." Modifications were made to approach this problem. First, the main portion of the task that asked participants to type the number of widgets they would make was changed from "*How many widgets would you make in a day if you had to create X widgets to earn a dollar*" to "*Considering what you can buy, how many widgets do you think you would make in a day if you had to create 1 widget to earn a dollar*?" The purpose of this change was to have participants attend to the menu, and frame it based on their opinion, rather than treating it like a math problem.

Other participants treated the task of making widgets as mandatory. This was especially apparent in pilot 1, where one participant responded the following way:

"*I wasn't sure if I had other means of making money in this scenario. I kinda imagined myself in some dystopian nightmare of only being able to earn money by making widgets all day, only being able to purchase the items listed. I eventually chose to die by refusing to work rather than continue living such an existence. If I knew I had other options to make money in this scenario,*

that would've drastically changed my answers. I would've stopped working for such poor pay much earlier."

While this response was initially humous, it provided a potential cue that the economy being described in the HPT was much more closed than what would be encountered by an individual receiving a token economy as a treatment. After Pilot 1 it was decided to frame the task as the participant completing volunteer work rather than their job being a "widget builder", with their requirement to tell the study team how hard they would want to work. They were told that they could stop working at any time and go home. No similar qualitative responses occurred in the main study.

Limitations

Despite MTurk allowing for participants to answer 12 HPTs in about one hour, conducting this study using a within-subject design put a strain on MTurk workers. This was best described by one participant who stated "*well, this was 2x as long as it should have been to keep someone's attention focused.*" To control for the possibility of experimental fatigue confounding results, participants were given all 12 tasks in a random order.

While the study design allowed for within subject comparisons, a between subject or mixed design may be more appropriate when using MTurk. This is primarily due to the large amount of non-systematic data and outliers that must be removed. According to Stein et al. (2015) there are several reasons why non-systematic data can be a frequent concern when using the hypothetical purchase task, including not paying attention to the prices on the task, typing incorrect responses, or failing to understand task directions. Of the 835 total hypothetical purchase tasks, the results of 256 (30.6%) were nonsystematic. There are several reasons why

nonsystematic data may have been prevalent in this study. The first reason is that participants may have either used the same value across all token production schedules (e.g. 1s for all 1-800 response requirements) or matched the response requirement in their response (1 for FR-1; 800 for FR-800). Both responses would have been eliminated using the Stein et al. (2015) method and were primarily due to participants completing the task quickly to maximize earnings. The second reason why non-systematic was observed, and further outliers were removed, was due to very broad response limits allowed in the hypothetical purchase task. In pilots 1 & 2, there was no limit to the values that participants could enter in the hypothetical purchase tasks, despite it being impossible to produce more than 1440 widgets in a 24-hour span. A 1440 widget limit was added in the main study. However, this still allowed for a very large range of responses to occur in each task. Future research should impose a more realistic limit on widget production. For example, a 480-widget production limit (1 per minute for 8 hours) may be more appropriate and control for outliers that are still included after the Stein et al. (2015) method was implemented. The third reason why systematic data was prevalent in the current study was the additional instructions beyond that of a typical HPT. This study was the first HPT to include a menu of purchasing options in addition to a more in-depth scenario.

Below is an overview of the same study procedures being implemented using a mixed model with both a between and within subject measure to help control for the large variability found in our dataset:

Type of Items on the Menu	**Number of Items on the Menu**		
Primary Reinforcer Only	3	6	12
Tangible Reinforcers Only	3	6	12

Activity Reinforcers Only	3	6	12
Mixed Menu	3	6	12

In this design, there would be four groups, each of which receive three hypothetical purchase tasks for each menu type. The within subject variable would be the number of items on the menu while the between subject variable would be the menu categories.

Finally, a completely between subjects design would appear similar. However, there would be 12 randomly assigned groups, each receiving a different level of the independent variables. Both methods would require an increase in the number of subjects, with the group design requiring a structured power analysis, which isn't prevalent in within subject designs.

Low R Squared Values

One significant limitation to this study was the amount of overall variation that was explained when applying the exponential model of demand to the data set. In the present study, only the six Activity model reached an R^2 value of over .50 (R^2 = .53). However, regression models that are applied to clinical research studies using human participants often produce lower R^2 values (Hamilton et al., 2015). This may also be true when regression models are applied to within-subject designs, that often have a smaller participant pool. Larger between subject designs that involve the use of a hypothetical purchase task often have larger goodness of fits. For example, in Roma et al. (2015), 1219 data sets were collected for participants who completed a hypothetical purchase task, R_2 values ranged from .83 to 1.0 (M = .98). To improve goodness of fit, future studies could employ a between-subjects design with a larger sample size. Additionally, narrowing the constraints on responding in the hypothetical purchase task could reduce the amount of variation in the data set, which would increase the R^2 value.

Idiosyncrasies of Menu Preferences

Another significant limitation to the present study was that participants did not complete any preference assessments prior to completing the hypothetical purchase task. Therefore, the menu was not created based on their preferences, which isn't typical in an applied setting. This could have had a suppressive effect on demand. Attempts were made to keep the descriptions of some of the menu items as general as possible. For example, rather than including a cheeseburger on the menu, an 8-inch sub, burger, or similarly prepared sandwich was included as a single menu item. While the HPTs were designed to account for variance preferences, the probability of a participant not preferring a menu item increased substantially as the menu size decreased. Additionally, participants did not state the items they were purchasing on the menu. Participants were also not asked in any qualitative questions if there were certain menu items that they were hypothetically purchasing.

Menu Prices

While menu prices were held constant throughout all study conditions, it was difficult to keep them equal for each menu category without deviating away from market value. This was particularly a problem in the primary reinforcer condition. Overall, food and drink items are less expensive than tangible and activity reinforcers. While prices of the primary reinforcer items were slightly inflated to reduce the price gap between primary reinforcers and the other categories, the most expensive primary reinforcer menu item was a personal pizza ($11). This is contrasted with the most expensive tangible item (a video game) which was priced at $30, and the most expensive activity (tickets to a sporting event), which was priced at $50. While decreasing the price of the activity and menu items on the menu would have closed the gap

between menu categories, this would have resulted in participants potentially seeking out discounted items on the menu, which may have confounded results.

Future Directions

To validate the application of Hypothetical Purchase tasks for measuring response output, future research should be conducted to compare hypothetical responses to actual earning and spending of tokens in a laboratory or applied setting. This would be modeled after similar hypothetical purchase task research, like Amlung et al. (2012) who compared responding during a hypothetical alcohol purchase task to a task with actual alcohol rewards and found a close correspondence between the two for both demand for and consumption of alcohol. Similar research was also done with real and hypothetical cigarette consumption (Wilson et al., 2016).

Most token economy research involves holding the token exchange schedule constant. The token exchange schedule (i.e. price) was held constant across all conditions in the present study. In other token research areas, such as token accumulation, the token exchange schedule is held constant at FR-1 when manipulating token production and exchange production schedules (Yankelevitz et al., 2008). In another study on token accumulation Regnier, Van Zandt, & DeFulio (2020) manipulated the cost of the items available in the marketplace to be relative to the price paid for them in a real-world setting. Future research using the task employed in the present study could parametrically assess the effects of generality and menu price on the production of widgets.

Another important token component schedule that requires more experimentation as it relates to demand is the exchange production schedule. Exchange production schedule increases may slightly decrease essential value. However, demand may be less elastic with increases in

exchange production than token production, which sharply decreased the production of widgets in this study. By requiring participants to accumulation tokens prior to exchanging them, the magnitude of back-up reinforcement available at the time of the exchange increases. Additionally, given that tokens bridge the gap between a target response and terminal reinforcement, increasing response requirements to make the exchange after a token has been earned should have less of an effect on responding than increasing response requirements to earn a token (Hackenberg, 2018). With increases in exchange production schedule, demand should become more elastic at higher FR schedule requirements. However, these predictions have yet to be tested empirically.

An additional behavior that is relevant to behavioral economics is token accumulation, defined as the conditions under which an individual will save money rather than spend it immediately (Hackenberg, 2018; Yankelevitz & Hackenberg, 2009). Token accumulation research has significant applied value and may serve as an indicator of performance. For example, in one study on the use of token systems to promote appropriate behavior, participants who save their tokens show performance decline over time (Winkler, 1973). From an applied perspective, designing a token economy to promote spending rather than saving may improve performance. However, the results of this research vary, which may indicate other moderating variables that affect the relationship between accumulation and performance. In Subramaniam et al. (2017), for example, participants who held a higher balance during a therapeutic workplace intervention for adherence to naltrexone also tended to have higher rates of heroin and cocaine abstinence. There has been a growing body of literature on token accumulation and the manipulation of the token component schedules to promote spending. In summary, increasing the token production schedule tends to decrease accumulation (Yankelevitz et al., 2008), while

increasing the exchange production schedule increases accumulation (Yankelevitz et al., 2008; Killeen, 1974; Regnier et al., 2020).

There have also been preliminary investigations of the effects of token generalizability on accumulation suggesting that increasing the number of menu options available increases accumulation. However, results varied, and further investigation is required (Regnier et al., 2020). Applied accumulation research has focused on the effects of accumulated reinforcers that are provided after a delay, compared to immediate, distributed reinforcers. DeLeon et al. (2014) found that task completion was highest when participants were given access to accumulated reinforcers contingent on larger fixed ratio schedules rather than shorter access of reinforcers contingent upon a low response requirement. Participants completed more tasks and preferred the accumulated, delayed reinforcers. Target behaviors often occur at lower levels in accumulated reinforcement conditions (Fulton et al., 2020; Robinson & Peter, 2019) and participants have more success in skill acquisition programs (Frank- Crawford et al., 2019).

While there is a growing body of research assessing the individual effects of token production schedules, exchange production schedules, and token generality on accumulation, the extent to which those variables interact has yet to be explored. A parametric analysis of token production schedule, exchange production schedule; and token generality's effect on human accumulation is warranted. Yankelevitz et al. (2008) ran a similar study with pigeons, with the exclusion of token generality manipulations. As was further displayed in the present study, token generality is a complex variable that may also serve to moderate the relationship between significant token variables, generality, and reinforcer value.

Conclusion

When designing token economies, little consideration is made regarding the back-up reinforcers included on the menu. If making menu decisions based on previous empirical evidence a therapist may consider a highly generalized menu. However, it wouldn't be entirely clear whether that involves a larger menu, and/or including many different types of back-up reinforcers. The present study was the first application of the hypothetical purchase task on demand for tokens and extends prior research on generalized token reinforcement by demonstrating the interaction between reinforcer categories and number of available menu items on demand. These results have important implications for the development and modification of token economies in behavior therapy and can assist a therapist to produce the most robust token economy possible.

References

Allen, K. E., Hart, B., Buell, J. S., Harris, F. T., & Wolf, M. M. (1964). Effects of social reinforcement on isolate behavior of a nursery school child. *Child Development, 35*(2), 511-518.

Amlung, M. T., Acker, J., Stojek, M. K., Murphy, J. G., & MacKillop, J. (2012). Is talk "cheap"? an initial investigation of the equivalence of alcohol purchase task performance for hypothetical and actual rewards. *Alcoholism: Clinical and Experimental Research, 36*(4), 716-724.

Ayllon, T., & Azrin, N. (1968). *The token economy: A motivational system for therapy and rehabilitation* Appleton-Century-Crofts, East Norwalk, CT.

Broadbent, J., & Dakki, M. A. (2015). How much is too much to pay for internet access? A behavioral economic analysis of internet use. *Cyberpsychology, Behavior, and Social Networking, 18*(8), 457–461.

Bugelski, R. (1938). Extinction with and without sub-goal reinforcement. *Journal of Comparative Psychology,* 26(1), 121–134.

Bullock, C. E., & Hackenberg, T. D. (2006). Second-order schedules of token reinforcement with pigeons: Implications for unit price. *Journal of the Experimental Analysis of Behavior, 85*(1), 95–106.

Burchard, J. D., & Barrera, F. (1972). An analysis of timeout and response cost in a programmed environment. *Journal of Applied Behavior Analysis, 5*(3), 271–282

Byrd, L. D. (1971). Responding in the pigeon under chained schedules of food presentation: The repetition of a stimulus during alternate components. *Journal of the Experimental Analysis of Behavior, 16*(1), 31–38.

Charnov, E. (1976). Optimal foraging: The marginal value theorem. *Theoretical Population Biology, 9,* 129–136.

Crowell, C. R., Anderson, D. C., Abel, D. M., & Sergio, J. P. (1988). Task clarification, performance feedback, and social praise: Procedures for improving the customer service of bank tellers. *Journal of Applied Behavior Analysis,* 21(1), 65–71.

Davis, D. R., Kurti, A. N., Skelly, J. M., Redner, R., White, T. J., & Higgins, S. T. (2016). A review of the literature on contingency management in the treatment of substance use disorders, 2009–2014. *Preventive Medicine,* 92, 36–46.

DeLeon, I. G., Chase, J. A., Frank Crawford, M.A., Carreau Webster, A. B., Triggs, M. M., Bullock, C. E., & Jennett, H. K. (2014). Distributed and accumulated reinforcement arrangements: Evaluations of efficacy and preference. *Journal of Applied Behavior Analysis, 47*(2), 293-313.

DeLeon, I. G., & Iwata, B. A. (1996). Evaluation of a multiple-stimulus presentation format for assessing reinforcer preferences. *Journal of applied behavior analysis, 29*(4), 519–533.

Donaldson, J. M., DeLeon, I. G., Fisher, A. B., & Kahng, S. (2014). Effects of and preference for conditions of token earn versus token loss. *Journal of Applied Behavior Analysis* , **47**, 537– 548.

Epstein, L. H., Dearing, K. K., & Roba, L. G. (2010). A questionnaire approach to measuring the relative reinforcing efficacy of snack foods. *Eating Behaviors, 11*(2), 67–73.

Ferster, C. B., & DeMyer, M. K. (1962). A method for the experimental analysis of the behavior of autistic children. *American Journal of Orthopsychiatry,* 32(1), 89–98.

Fantuzzo, J. W., Rohrbeck, C. A., Hightower, A. D., & Work, W. C. (1991). Teachers' use and children's preferences of rewards in elementary school. *Psychology in the Schools, 28,* 175–181.

Fisher, W., Piazza, C. C., Bowman, L. G., Hagopian, L. P., Owens, J. C., & Slevin, I. (1992). A comparison of two approaches for identifying reinforcers for persons with severe and profound disabilities. *Journal of Applied Behavior Analysis,* 25, 491–498.

Frank-Crawford, M., Borrero, J. C., Newcomb, E. T., Chen, T., & Schmidt, J. D. (2019). Preference for and efficacy of accumulated and distributed response– reinforcer arrangements during skill acquisition. *Journal of Behavioral Education*, *28*(2), 227-257.

Fulton, C. J., Tiger, J. H., Meitzen, H. M., & Effertz, H. M. (2020). A comparison of accumulated and distributed reinforcement periods with children exhibiting escape-maintained problem behavior. *Journal of Applied Behavior Analysis*, *53*(2), 782-795.

Gable, R. A., Hester, P. H., Rock, M. L., & Hughes, K. G. (2009). Back to Basics: Rules, Praise, Ignoring, and Reprimands Revisited. *Intervention in School and Clinic*, 44(4), 195–205.

Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W., & Smith, G. D. (2006). Indicators of socioeconomic position (part 1). *Journal of Epidemiology and Community Health, 60*(1), 7–12.

Hackenberg, T. D. (2009). Token reinforcement: A review and analysis. *Journal of the Experimental Analysis of Behavior, 91*(2), 257–286.

Hackenberg, T. D. (2018). Token reinforcement: Translational research and application. *Journal of Applied Behavior Analysis, 51*(2), 393–435.

Hamilton, D. F., Ghert, M., & Simpson, A. H. (2015). Interpreting regression models in clinical outcome studies. *Bone & joint research*, *4*(9), 152–153.

Hodos, W. (1961). Progressive ratio as a measure of reward strength. *Science, 134*, 943–944.

Hyde, T. S. (1976). The effect of pavlovian stimuli on the acquisition of a new response. *Learning and Motivation*, 7(2), 223–239.

Ivy, J. W., Meindl, J. N., Overley, E., & Robson, K. M. (2017). Token economy: A systematic review of procedural descriptions. *Behavior Modification, 41*(5), 708–737.

Iwata, B. A., & Bailey, J. S. (1974). Reward versus cost token systems: An analysis of the effects on students and teacher. *Journal of Applied Behavior Analysis* , **7**, 567– 576.

Jacobs, E. A., & Bickel, W. K. (1999). Modeling drug consumption in the clinic using simulation procedures: demand for heroin and cigarettes in opioid-dependent outpatients. *Experimental and clinical psychopharmacology, 7*(4), 412–426.

Jarmolowicz, D. P., Reed, D. D., Francisco, A. J., Bruce, J. M., Lemley, S. M., & Bruce, A. S.. (2018). Modeling effects of risk and social distance on vaccination choice. *Journal of the Experimental Analysis of Behavior, 110*, 39–53.

Jowett Hirst, E. S., Dozier, C. L., & Payne, S. W. (2016). Efficacy of and preference for reinforcement and response cost in token economies. *Journal of Applied Behavior Analysis, 49*(2), 329–345.

Kahng, S., Boscoe, J. H., & Byrne, S. (2003). The use of an escape contingency and a token economy to increase food acceptance. *Journal of Applied Behavior Analysis, 36,* 349–353.

Kelleher, R. T. (1966). Conditioned reinforcement in second-order schedules. *Journal of the Experimental Analysis of Behavior, 9*(5), 475–485.

Kelleher, R. T., & Gollub, L. R. (1962). A review of positive conditioned reinforcement. *Journal of the Experimental Analysis of Behavior, 5*(4), 543–597.

Kiselica, A. M., Webber, T. A., & Bornovalova, M. A. (2016). Validity of the alcohol purchase task: A meta-analysis. *Addiction, 111*(5), 806–816.

Koffarnus, M. N., Franck, C. T., Stein, J. S., & Bickel, W. K. (2015). A modified exponential behavioral economic demand model to better describe consumption data. *Experimental and clinical psychopharmacology, 23*(6), 504–512.

Leon, Y., Borrero, J. C., & DeLeon, I. G. (2016). Parametric analysis of delayed primary and conditioned reinforcers. *Journal of Applied Behavior Analysis, 49*(3), 639–655.

Kranak, M. P., Alber-Morgan, S., & Sawyer, M. R. (2017). A parametric analysis of specific praise rates on the on-task behavior of elementary students with autism. *Education and Training in Autism and Developmental Disabilities*, 52(4), 453–464.

Jones, R.T. & Kazdin, A. E. (1975). Programming response maintenance after withdrawing token reinforcement. *Behavior Therapy, 6, 153–164.*

Madden, G. J., & Bickel, W. K. (Eds.). (2010). *Impulsivity: The behavioral and neurological science of discounting.* American Psychological Association.

Malesza, M. (2019). The effects of potentially real and hypothetical rewards on effort discounting in a student sample. *Personality and Individual Differences, 151,* Article 108807

Mitchell, S. H. (2004). Effects of short-term nicotine deprivation on decision-making: Delay, uncertainty and effort discounting. *Nicotine & Tobacco Research, 6*(5), 819-828.

Nastasi, J. A., Sheppard, R. D., & Raiff, B. R. (2020). Token-economy-based contingency management increases daily steps in adults with developmental disabilities. *Behavioral Interventions, 35*(2), 315–324.

O'Donnell, J., Crosbie, J., Williams, D. C., & Saunders, K. J. (2000). Stimulus control and generalization of point-loss punishment with humans. Journal of the Experimental Analysis of Behavior, 73(3), 261–274.

O'Neill, R. E., Horner, R. H., Albin, R. W., Sprague, J. R., Storey, K., & Newton, J. S. (1997). *Functional assessment and program development for problem behavior: A practical handbook*. Pacific Grove, CA.

Pietras, C. J., & Hackenberg, T. D. (2005). Response-cost punishment via token loss with pigeons. *Behavioural Processes*, 69(3), 343-356.

Raiff, B. R., Bullock, C. E., & Hackenberg, T. D. (2008). Response-cost punishment with pigeons: Further evidence of response suppression via token loss. *Learning & Behavior, 36*, 29–41.

Reed, D. D., Partington, S. W., Kaplan, B. A., Roma, P. G., & Hursh, S. R. (2013). Behavioral economic analysis of demand for fuel in north america. *Journal of Applied Behavior Analysis, 46*(3), 651–655.

Regnier, S., Van Zandt, N., & DeFulio, A. 2020. An exploratory analysis of human token accumulation. *Experimental Analysis of Human Behavior Bulletin, 32, 32-39.*

Rescorla, R. A. (1968). Probability of shock in the presence and absence of cs in fear conditioning. *Journal of Comparative and Physiological Psychology*, 66(1), 1–5.

Robinson, N., & St Peter, C.,C. (2019). Accumulated reinforcers increase academic responding and suppress problem behavior for students with Attention Deficit hyperactivity disorder. *Journal of Applied Behavior Analysis*, *52*(4), 1076-1088.

Sheffer, C. E., Bickel, W. K., Franck, C. T., Panissidi, L., Pittman, J. C., Stayna, H., & Evans, S. (2017). Improving tobacco dependence treatment outcomes for smokers of lower socioeconomic status: A randomized clinical trial. *Drug and Alcohol Dependence*, *181*, 177–185.

Soares, D. A., Harrison, J. R., Vannest, K. J., & McClelland, S. S. (2016). Effect size for token economy use in contemporary classroom settings: A meta-analysis of single-case research. *School Psychology Review*, *45*(4), 379–399.

Stein, J. S., Koffarnus, M. N., Snider, S. E., Quisenberry, A. J., & Bickel, W. K. (2015). Identification and management of nonsystematic purchase task data: Toward best practice. *Experimental and clinical psychopharmacology*, *23*(5), 377–386.

Stevens, C., Sidener, T. M., Reeve, S. A., & Sidener, D. W. (2011). Effects of behavior-specific and general praise, on acquisition of tacts in children with pervasive developmental disorders. *Research in Autism Spectrum Disorders*, *5*(1), 666–669.

Strickland, J. C., Marks, K. R., & Bolin, B. L. (2020). The condom purchase task: A hypothetical demand method for evaluating sexual health decision-making. *Journal of the Experimental Analysis of Behavior, 113*(2), 435–448.

Subramaniam, S., DeFulio, A., Jarvis, B. P., Holtyn, A. F., & Silverman, K. (2017). Earning, Spending, and Drug Use in a Therapeutic Workplace. *The Psychological record*, *67*(2), 273–283.

Tan, L., & Hackenberg, T. D. (2015). Pigeons' demand and preference for specific and generalized conditioned reinforcers in a token economy. *Journal of the Experimental Analysis of Behavior, 104*(3), 296–314.

Waddell, T. R., Leander, J. D., Webbe, F. M., & Malagodi, E. F. (1972). Schedule interactions in second-order fixed-interval (fixed-ratio) schedules of token reinforcement. *Learning and Motivation, 3*(1), 91–100.

Webbe, F. M., & Malagodi, E. F. (1978). Second-order schedules of token reinforcement: Comparisons of performance under fixed-ratio and variable-ratio exchange schedules. *Journal of the Experimental Analysis of Behavior, 30,* 219–224.

Weinstock, J., Mulhauser, K., Oremus, E. G., & D'Agostino, A. R. (2016). Demand for gambling: Development and assessment of a gambling purchase task. *International Gambling Studies, 16,* 316–327.

Williams, W. A., & Fantino, E. (1994). Delay reduction and optimal foraging: Variable-ratio search in a foraging analogue. *Journal of the Experimental Analysis of Behavior, 61,* 465–477.

Wilson, A. G., Franck, C. T., Koffarnus, M. N., & Bickel, W. K. (2016). Behavioral economics of cigarette purchase tasks: Within-subject comparison of real, potentially real, and hypothetical cigarettes. *Nicotine & Tobacco Research, 18*(5), 524–530.

Windsor, J., Piche, L. M., & Locke, P. A. (1994). Preference testing: A comparison of two presentation methods. *Research in Developmental Disabilities, 15*, 439–455.

Wolfe, J. B. (1936). Effectiveness of token rewards for chimpanzees. *Comparative Psychology Monographs, 12*, 72.

Yankelevitz, R. L., Bullock, C. E., & Hackenberg, T. D. (2008). Reinforcer accumulation in a token reinforcement context with pigeons. *Journal of the Experimental Analysis of Behavior, 90*(3), 283-99.

Yu, J., Liu, L., Collins, R. L., Vincent, P. C., & Epstein, L. H. (2014). Analytical Problems and Suggestions in the Analysis of Behavioral Economic Demand Curves. *Multivariate behavioral research*, *49*(2), 178–192.

Zimmerman, E. H., & Zimmerman, J. (1962). The alteration of behavior in a special classroom situation. *Journal of the Experimental Analysis of Behavior*, 5(1), 59–60.

Appendix A

Pilot Menu

Item	Cost
A single serving of chips, chocolate, or candy	$2
Hot coffee, tea, or soft drink	$2
Glass of beer or wine	$4
8-inch sub, burger, or similar prepared food	$6
Veggie salad for one	$6
1-month access to internet subscription service (e.g., video services, news services)	$8
30 minutes of gym time	$8
Movie pass	$10
Board or card game	$16
A book	$16
Tank of gas (15 gal)	$24
A video game	$48

Appendix B

12 Item Menus for the Main Study

12 Item Tangible Menu

Item	Cost
1 movie pass	$10
Make-up or other beauty product	$8
A video game	$30
Board or card game	$16
Piece of wall art ($10 value)	$10
Preferred article of clothing (shirt, pants, etc.)	$15
Tank of Gas (15 gal)	$24
A candle (scent of your choosing)	$10
24-pack of sports cards	$15
Set of headphones	$20
Bucket of golf-balls at the driving range	$8
A book	$10

12 Item Mixed Menu

Item	Cost
8-inch sub, burger, or similarly prepared food	$8
30 minutes on your smart phone	$6
Tank of gas (15 gal)	$24
Hot coffee or tea	$4
Going to a professional sporting event (value $50)	$50
Make-up or other beauty product	$8
Veggie salad for one	$6
1-month access to internet subscription service (e.g., video services, news services)	$8
A preferred article of clothing (shirt, pants, etc.)	$15
A soft drink (12 oz)	$3
30 minutes of gym time	$8
A video game	$30

12 Item Primary Reinforcer Menu

Item	Cost
Single serving of chips	$2
Bottled water	$2
8-inch sub, burger, or similarly prepared food	$8
Veggie salad for one	$6
Glass of beer or wine	$6
Pint of ice cream	$5
Personal Pizza	$11
Soft Drink (12 oz)	$3
Hot Coffee or Tea	$4
Single Serving of Chocolate	$3
Single Serving of Candy	$3
Small fruit salad	$5

12 Item Activity Menu

Item	Cost
1-month access to internet subscription service (e.g., video services, news services)	$8
An hour hike with a friend	$10
30 access to music shop	$15
30 minutes of gym time	$8
30 minutes of time playing preferred video game	$8
Trip to a museum	$15
Going to a professional sporting event ($50 value)	$50
A one-hour cooking class	$40
30 minutes on your smart phone	$6
30 minutes of time on social media	$4
30 minutes of arts and crafts	$6
Trip to an arcade	$10

Appendix C

6 Item Menus

6 Item Tangible Menu

Item	Cost
1 movie pass	$10
Make-up or other beauty product	$8
A video game	$30
Board or card game	$16
Piece of wall art ($15 value)	$15
Preferred article of clothing (shirt, pants, etc.)	$15

6 Item Mixed Menu

Item	Cost
8-inch sub, burger, or similarly prepared food	$8
30 minutes on your smart phone	$6
Tank of gas (15 gal)	$24
Hot coffee or tea	$4
Going to a professional sporting event (value $50)	$50
Make-up or other beauty product	$8

6 Item Primary Reinforcer Menu

Item	Cost
Single serving of chips	$2
Bottled water	$2
8-inch sub, burger, or similarly prepared food	$8
Veggie salad for one	$6
Glass of beer or wine	$6
Pint of ice cream	$5

6 Item Activity Menu

Item	Cost
1-month access to internet subscription service (e.g., video services, news services)	$8
An hour hike with a friend	$15
30 minutes access to music shop	$6
30 minutes of gym time	$8
30 minutes of time playing preferred video game	$8
Trip to a museum	$15

Appendix D

3 Item Menus

3 Item Tangible Menu

Item	Cost
1 movie pass	$10
Make-up or other beauty product	$8
A video game	$30

3 Item Mixed Menu

Item	Cost
8-inch sub, burger, or similarly prepared food	$8
30 minutes on your smart phone	$6
Tank of gas (15 gal)	$24

3 Item Primary Reinforcer Menu

Item	Cost
A single serving of chips	$2
A bottled water	$2
8-inch sub, burger, or similarly prepared food	$8

3 Item Activity Menu

Item	Cost
1-month access to internet subscription service (e.g., video services, news services)	$8
An hour hike with a friend	$10
30 minutes access to music shop	$15

Appendix E

Comprehension Test Screener

1) If participant scored 100% on the comprehension test: *Great job! You passed the screener. It is time for the main part of the study. In this part of the study you will imagine that you have volunteered your time to participate in a project in which you are building widgets. You earn tokens or money that can only be used at the projects market. There are no right or wrong answers in this part of the survey. When answering the questions, the menu will change, so consider the menu of items available and think about how hard you would want to work. You could stop working at any time and go home.*

2) If participant failed the screener: *You did not pass the screener questions. When you press next, you will be exited from the survey*

Appendix F

Study Break Periods

1) Eligibility Screener Break: *Prior to beginning the study, you will be required to complete several training questions, followed by a brief comprehension test. If you do not pass the comprehension test, you will be excluded from the study and ineligible for compensation. Please pay careful attention to the following examples and questions. This section should take less than 3 minutes. If you exit out of this survey, you will not be able to reopen it.*

2) Comprehension Test Break: *Now it is time for the comprehension test. Please remember that if you do not pass the comprehension test, you will be excluded from the study and ineligible for compensation*

3) Demographics Break: *We will now ask you some basic demographic questions. Click to continue.*

Appendix G

Open Ended and Attention Checks

Opened Ended: *Please comment on your thought processes when answering the hypothetical questions.*

Attention Check #1: *What year is it?*

- 1991
- 1954
- 2018
- 2021

Attention Check #2: Would you rather have **$1000 immediately** or have **$1 in a year?**

$1000 immediately $1 in a year

Attention Check #3: *Which of the following is **NOT** a state in the United States?*

- *Massachusetts*
- *Michigan*
- *Beijing*
- *Texas*

Appendix H

Completion Code Question

Your completion ID is ${e://Field/Success}, make sure you save this number as this indicates you have completed the survey.

Please paste this number ${e://Field/Success} into MTurk before pressing next. If you press next without saving the number, you may not be compensated for your participation.

After submitting this code into MTurk, you will be compensated $10 for completing the survey in it's entirety.

When you press next, you will have completed the survey.

Appendix I

Informed Consent for Main Study

WESTERN MICHIGAN UNIVERSITY
IRB Approved
Approved for use for one year from this date:
DEC 10 2020
WMU IRB Office

Western Michigan University
Department of Psychology

Please read this consent information before you begin the survey.

Principal Investigator: Anthony DeFulio
Student Investigator: Sean Regnier

You are invited to participate in this research project titled "The Effects of Menu Manipulations on Token Demand: An experimental Survey"

STUDY SUMMARY: This consent form is part of an informed consent process for a research study and it will provide information that will help you decide whether you want to take part in this study. The purpose of the research is to study how demand for tokens and money changes as the effort required to earn them increases. When you begin the survey, you are consenting to participate in the study. If you do not agree to participate in this research project, simply exit now. If, after beginning the survey, you decide that you do not wish to continue, you may stop at any time. You may choose to not answer any question for any reason.

It will serve as Sean Regnier's dissertation for the requirements of a PhD. If you take part in the research, you will be asked to answer hypothetical scenarios where you are asked how much you would work when the amount of payment you receive for your work is increased or decreased.

Your replies will be completely anonymous, so do not put your name anywhere on the survey. Your time in the study will take between 30-60 minutes. Possible risks and costs to you for taking part in the study are that you may become bored while answering questions. There is a risk that the confidential data collected in this study could be revealed to someone outside of this study such as a friend, relative, or outside organization. This could occur due to breaches in the security of data storage, such as a computer hacker breaking into the survey platform. To minimize the risk of confidential information being revealed, all data collected will be password protected and kept on a secure server accessible only be study staff. You are also only identifiable by your Worker ID. There are no direct benefits to you for participating in this study. Your alternative to taking part in the research study is not to take part in it.

The de-identified (anonymous) information collected for this research may be used by or distributed to investigators for other research without obtaining informed consent from you.

WESTERN MICHIGAN UNIVERSITY

IRB Approved

Approved for use for one year from this date:

DEC 10 2020

Amy Naugle

WMU IRB Office

Should you have any questions prior to or during the study, you can contact the principal investigator, Anthony DeFulio at 269-387-4459 or anthony.defulio@wmich.edu or the student investigator, Sean Regnier at 413-427-7489 or by email at sean.d.regnier@wmich.edu. You may also contact the Chair, Institutional Review Board at 269-387-8293 or the Vice President for Research at 269-387-8298.

This consent document has been approved for use for one year by the Western Michigan University Institutional Review Board (WMU IRB) as indicated by the stamped date and signature of the board chair in the upper right corner.

By accepting and submitting this HIT, you agree to these terms.

Appendix J

WMU IRB Approval

WESTERN MICHIGAN UNIVERSITY

Human Subjects Institutional Review Board

Date: December 10, 2020

To: Anthony DeFulio, Principal Investigator
Sean Regnier, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: IRB Project Number 20-12-11

This letter will serve as confirmation that your research project titled "The Effects of Token Menu Manipulations on Token Demand: An Experimental Survey" has been **approved** under the **exempt** category of review by the Western Michigan University Institutional Review Board (IRB). The conditions and duration of this approval are specified in the policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may **only** be conducted exactly in the form it was approved.
You must seek specific board approval for any changes to this project (e.g., ***add an investigator, increase number of subjects beyond the number stated in your application, etc.***). Failure to obtain approval for changes will result in a protocol deviation.

In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the IRB for consultation.

The Board wishes you success in the pursuit of your research goals.

A status report is required on or prior to (no more than 30 days) December 9, 2021 and each year thereafter until closing of the study. The IRB will send a request.

When this study closes, submit the required Final Report found at https://wmich.edu/research/forms.

Note: All research data must be kept in a secure location on the WMU campus for at least three (3) years after the study closes.

251 W. Walwood Hall, Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293, FAX: (269) 387-8276