A Behavioral Economic Evaluation of Preference and Reinforcer Assessment Methods in Organizational Behavior Management

By
© 2020
Matthew D. Novak

Submitted to the graduate degree program in the Department of Applied Behavioral Science and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson: Florence D. DiGennaro Reed, Ph.D., BCBA-D

Claudia L. Dozier, Ph.D., BCBA-D

Edward K. Morris, Ph.D.

Thomas Zane, Ph.D., BCBA-D

Jacob T. Fowles, Ph.D.

Date Defended: 28 April 2020

The dissertation committee for Matthew D. Novak
certifies that this is the approved version of the following dissertation:

A Behavioral Economic Evaluation of Preference and Reinforcer Assessment Methods in Organizational Behavior Management

Chairperson: Florence D. DiGennaro Reed, Ph.D., BCBA-D

Date Approved: 28 April 2020

Abstract

There are several available methods for identifying incentives that employees would prefer to work for. However, few organizational behavior management studies have directly evaluated the reliability and validity of these formats. This experiment was a translational research study evaluating the correspondence between and the validity of three preference assessment formats: a Likert-type *survey*, a *ranking* task, and a *hypothetical work task*, which asked participants whether they would be willing to complete a given number of work units to earn an incentive. Participants first completed the three preference assessments, followed by two multielement reinforcer assessments. The first reinforcer assessment examined responding for each incentive at a low, fixed response requirement. In the second reinforcer assessment, participants worked for each incentive on a progressive ratio reinforcement schedule. Finally, participants completed follow-up preference assessments. There was high correspondence across preference assessment formats and between preference and reinforcer assessment outcomes. Results demonstrated the viability of this methodology for evaluations of preference and reinforcer assessments in work-related contexts.

Keywords: organizational behavior management, behavioral economics, incentives, preference assessment, reinforcer assessment

Acknowledgements

This project is a culmination of efforts from a great many people. First, thank to my committee members, Drs. Edward Morris, Thomas Zane, and Jacob Fowles, for your guidance and feedback throughout this process as well as your flexibility during this period of tumult. A special thank you goes to Drs. Claudia Dozier and Derek Reed, not only for your service on my committee but also the many different roles that you have played in my development over the past five years. Thank you for the many ways that you have each been, at times, a second advisor to me, and thank you for playing such critical roles in expanding both the breadth and depth of my research and professional skills.

A very special thank you goes to Dr. Florence DiGennaro Reed. There is simply no way to adequately thank you for all you have done to foster so much growth in these five years. Your contributions in any of your roles as advisor, committee chair, lab director, and department chair alone would warrant special mention here, and even then, it would be nearly impossible to summarize my gratitude in a single paragraph. Thank you for arranging an environment that, at every level, maximized the opportunity for growth and thank you for always being available to provide guidance in all areas of life. I will be forever grateful for this opportunity and for your efforts as both a mentor and colleague along the way.

Thank you to the many people who make up my environment. To the members to the Applied Behavioral Science community at large, thank you for maintaining a culture that supports intellectual and professional growth. I deeply appreciate the opportunity to be a part of this department's rich history. To members of the Performance Management and Applied Behavioral

Economics Labs, past and present, thank you for your constant and unwavering support. Your daily efforts and interactions played a critical role at each step in this project and my experience at KU as a whole.

I owe the deepest thank you to Ashley Romero, Marcella Hangen, Abby Blackman, and Tyler Erath. Thank you for being such an integral part of all aspects of life, both at and outside of work. Until a couple of months ago, I would have been hard-pressed to think of many days over the past few years without your involvement—a feat which, quite honestly, is more deserving of a medal for your part, but a thank you will have to suffice. So, thank you for tolerating my own idiosyncrasies and for making every day better. Whether we were discussing research, grabbing drinks, rehearsing talks, watching KU games, cooking food, or really anything else, I have truly valued each of these moments. None of this would have been possible without you.

Finally, thank you to my mom and dad, Danielle, Willis, Avery, and Ralph for your never-ending support from the very beginning. Thank you for your constant flexibility over the past five years and for always supporting my goals.

Table of Contents

List of Illustrative Materials

Appendix Index

A Behavioral Economic Evaluation of Preference and Reinforcer Assessment Methods in Organizational Behavior Management

Applied behavior analysis is a scientific discipline that applies the science of behavior to produce socially significant behavior change (Baer et al., 1968). The emphasis on social significance has been a defining feature of the field dating back to its founding (Wolf, 1978), and one means toward achieving social significance is through applications in areas that directly affect a large portion of the population (Friman, 2010). Nearly two-thirds of all United States citizens take part in some form of employment (Bureau of Labor Statistics, 2019b), and among full-time employees, work is the single activity in which they spend the greatest proportion of their waking hours (Bureau of Labor Statistics, 2019a). Thus, the study of work-related behavior is an opportunity for both socially significant and far-reaching application.

Organizational behavior management (OBM) is a discipline of applied behavior analysis dedicated to the assessment and intervention of employee behavior to improve individual and group performance and effectiveness within industry, business, and government (Frederiksen & Lovett, 1980). Within OBM are three broad sub-disciplines: behavior-based safety, behavioral systems analysis, and performance management (Wilder et al., 2009). Behavior-based safety involves targeted behavior-analytic assessment and intervention to promote safe workplace environments and employee behavior (Boyce & Geller, 2001). Behavioral systems analysis is an approach to understanding organizational systems by viewing how interconnected components of an organization interact with and affect each other (Diener et al., 2009). Performance management involves the direct application of applied behavior analysis through assessment and intervention to alter antecedent and consequent events controlling employee behavior (Daniels & Bailey, 2014).

Performance management interventions may be broadly classified into two major concentrations: (a) training employees the skills necessary to perform job duties and (b) ongoing support to promote maintenance of those skills and improve overall performance (Novak et al., 2019; Reid & Parsons, 2000). A large body of staff training intervention research has demonstrated the effectiveness of performance- and competency-based training methods (Parsons et al., 2012), and this robust literature has consistently demonstrated a methodology for successfully training necessary skills. However, training is only one component of a broader package. It is important to arrange a workplace environment that promotes maintenance and improvement of trained skills over time. This second component, which can be referred to as coaching (DiGennaro Reed et al., 2013), often involves manipulations to the workplace environment to set the occasion for and reinforce desired performance. In a review of research articles published in OBM's flagship journal, the *Journal of Organizational Behavior Management,* VanStelle et al. (2012) found that the most researched interventions for maintaining performance were feedback, antecedent manipulations (e.g., task clarification, prompting), and incentives.

Incentives in OBM

Incentives are rewards that are delivered in the workplace contingent on a specific employee behavior or permanent product (Daniels & Bailey, 2014). They can take many forms, including money (e.g., Bucklin & Dickinson, 2001), tangible items (e.g., food; Kortick & O'Brien, 1996), activities and privileges (e.g., selecting job duties; Reed et al., 2012), or breaks from work (e.g., Wilder et al., 2006).

From a conceptual standpoint, incentives are putative reinforcers. They are delivered contingent on a specific behavior, with an assumption that their delivery will increase the future rate of that behavior. Because incentives may involve delivery of a preferred stimulus or removal of an aversive stimulus, they may be classified as either positive or negative (putative) reinforcers. In the extant literature, the term *incentive* refers to contingent rewards for desirable behavior, independent of whether the rewards actually produce increased rates of behavior (e.g., VanStelle et al., 2012), which is why the reinforcing nature of incentives is putative. Additionally, consequences that function as reinforcers for desirable work performance extend well beyond any programmed incentives that may be in place (e.g., positive interactions from coworkers, completion of a task). Thus, not all workplace reinforcers are incentives, nor do all incentives function as reinforcers.

Monetary Incentives

Monetary incentives are the most frequently studied form of incentive in OBM, appearing in nearly twice as many publications as nonmonetary incentives (VanStelle et al., 2012). Across several studies, monetary incentives have reliably produced increases in performance compared to noncontingent pay; these increases could be attributed to the existence of a contingency, not the incentive amount (Poling et al., 2000). One advantage of using money as an incentive is that employees can use it in any manner they prefer; as a generalized conditioned reinforcer, money has the advantage of being resistant to satiation and it is likely highly valued by all employees (Catania, 2013; Daniels & Bailey, 2014).

Despite the success of monetary incentives, several barriers may impede their implementation in practice. First, when considering the sustainability of an intervention (Sigurdsson & Austin, 2006), organizational leaders may not be able to make a long-term commitment to the resources necessary for a monetary incentive program. Second, monetary incentives may be subject to federal and state labor regulations (e.g., Fair Labor Standards Act of 1938), which may create logistical challenges for an organization.

In addition to the logistical challenges, limitations associated with the reinforcing properties of the incentive itself may hinder its use. First, monetary incentives would likely be delivered as additional funds in a biweekly paycheck, which introduces a delay between when the desirable performance occurs and when the incentive is received, thus weakening the contingency between the desired behavior and delivery of the incentive (Mazur, 1993). As a result, the incentive may be less noticeable and will likely be less effective. Second, although money is likely highly valued by all employees, other items may be even more preferred. For example, Wine et al. (2013) found that most employees favored at least one other type of incentive more than money (e.g., time off from work, opportunity to choose work location). Moreover, individuals often spend money for nonpreferred activities (e.g., medical bills), whereas nonmonetary incentives may be exclusively preferred items or activities (e.g., food).

Nonmonetary Incentives

Nonmonetary incentives provide a practical, cost-effective solution for supervisors who want to reinforce desired performance. However, one consideration with nonmonetary incentives is that individuals' preferences (i.e., what items they value the most) will vary widely from employee to employee. Research assessing the relative effectiveness of different forms of reinforcers has demonstrated that behavior-analytic interventions are more effective when high-preference (HP) items are used as reinforcers than when low-preference items (LP) are used (e.g., Piazza et al., 1996). Thus, the identification of HP incentives is likely an important factor in the success and long-term maintenance of OBM interventions.

Supervisors looking to identify employees' HP incentives may be inclined to make these predictions on their own. However, research examining supervisors' predictions of their employees' preferred incentives has demonstrated that these predictions are generally inconsistent with employee reports (Wilder et al., 2007; Wilder et al., 2011). Another approach to identifying employee preferences is to simply ask them which incentives they prefer. Daniels and Bailey (2014) outline three arguments against using this approach. First, employees may have difficulty thinking of a reasonable incentive, as they likely do not know what the available options are. Second, employees may not tell you what incentives they prefer. These first two arguments describe patterns of behavior that are likely the result of various environmental factors. For example, employees simply may not have much experience identifying incentives for their own work performance. The third argument is from a logistical standpoint: employees may request incentives that are too expensive to be sustainable or simply out of the organization's control. Taken together, these factors suggest that to facilitate employee decision making and account for logistical constraints, consideration should be put into the format in which preference is assessed.

Brief History of Preference Assessment Research

Preference assessments are systematic, objective methods used to identify an individual's preferred stimuli. The stimuli can include various tangible objects, foods, activities, or other environmental arrangements (Hagopian et al., 2004; Piazza et al., 2011)[1]. Much of the research to date involves methods where preference is assessed through direct observation of behavior (Kang et al., 2013; Virués-Ortega et al., 2014). Preference assessment research began in response to difficulty identifying therapeutic reinforcers for individuals who had limited verbal repertoires

[1] The range of stimulus classes (e.g., objects, food, activities, privileges, reinforcement-schedule arrangements) that can be included in a preference assessment will be collectively referred to as *items*.

(Fisher et al., 1992; Pace et al., 1985; Wacker et al., 1985), and the majority of preference assessment research over the past 35 years has been conducted in clinical and educational contexts with individuals with intellectual and developmental disabilities (Virués-Ortega et al., 2014). Virués-Ortega et al. (2014) conducted a systematic review of all studies using the term "preference assessment" and found that of 108 articles examining preference assessments, 78 were conducted with individuals with intellectual and developmental disabilities. The following section will provide a brief review of preference assessment methodology commonly used in clinical and educational settings—recent systematic reviews of preference assessment literature with individuals with intellectual and developmental disabilities that provide greater depth can be found in Kang et al. (2013) and Virués-Ortega et al.

The process of conducting a preference assessment begins by identifying an array of items. This is typically done through informal discussion, completion of an informant instrument (e.g., Fisher et al., 1996), or both. Items suggested to be likely reinforcers can then be included in the array for the preference assessment, which can take one of several formats (DeLeon et al., 2013). In a review of preference assessment studies with individuals with disabilities, Kang et al. (2013) reported that the two most accurate and commonly used methods were the paired stimulus (Fisher et al., 1992) and multiple-stimulus-without-replacement (MSWO; DeLeon & Iwata, 1996) formats. In the paired stimulus format, an individual is asked to select one of two simultaneously presented items and is given the selected item for a brief duration (e.g., until consumed, after a set passage of time). The process is repeated until all possible pairings are presented and items can then be ranked in order of the percentage of opportunities they were chosen, with HP items being those selected most often. The MSWO is a similar but briefer method in which all items in the array are presented to the individual simultaneously. As items are selected, they are removed from the array until no items remain. This process can be conducted one or more times and a preference hierarchy can be assigned based on the order in which items are chosen.

Reinforcer Efficacy

An underlying assumption in preference assessment use is that HP items are likely more efficacious reinforcers than LP items. However, the effort involved in a preference assessment is often very low (e.g., reaching out to touch an object; Kang et al., 2013; Virués-Ortega et al., 2014) relative to a desired behavior in a therapeutic context (e.g., completing an academic task). Thus, it is possible that HP items may not actually function as reinforcers for meaningful, desired behavior.

One common method used to test this assumption is to conduct a reinforcer assessment (Piazza et al., 2011). Reinforcer assessments involve a direct assessment of whether a given item functions as a reinforcer for a target behavior, which is typically a simple response (e.g., pressing a button; Roscoe et al., 1999). The process begins by establishing a baseline rate of behavior, with no contingencies on a targeted behavior. Next, the putative reinforcer is delivered contingent on the target behavior. If the individual’s rate of behavior increases during sessions with the reinforcement contingency in place, the item can be considered an efficacious reinforcer. Reinforcer assessment arrangements may also make an alternative response for which there are no programmed reinforcement contingencies (e.g., a second button or leisure activity) concurrently available to ensure that engagement with the target behavior is due to the reinforcement contingency and not because it is the sole activity available (DeRosa & Roane, 2015; Piazza et al., 1996).

Reinforcer assessments can be used to evaluate the construct validity of a preference assessment, or the extent to which a preference assessment adequately measures reinforcing efficacy (VandenBos, 2015). For example, Northup et al. (1996) conducted preference assessments in three different formats (two variants of the paired stimulus format and a Likert-type survey), each using the same five-item array, with four children. Items were then classified as HP or LP based on criteria specific to each assessment format. Following the preference assessments, Northup et al. conducted reinforcer assessments with each of the five items, plus an arbitrary control item. Items that maintained behavior above baseline and control rates were designated as showing clear reinforcement effects; thus, these items were considered efficacious reinforcers.

Northup et al. (1996) quantified construct validity by calculating each of the three preference assessment formats' accuracy in predicting reinforcer assessment outcomes. Specifically, the predictor variable was an assessment's HP–LP classification for each item and the outcome variable was whether the item was an efficacious or inefficacious reinforcer. With these variables are four possible outcomes: true positive, true negative, false positive, and false negative. True positives were HP items that were efficacious reinforcers, true negatives were LP items that were inefficacious reinforcers, false positives were HP items that were inefficacious reinforcers, and false negatives were LP items that were efficacious reinforcers. Northup et al. calculated accuracy by dividing the sum of true positives and true negatives by the sum of all positives and negatives and multiplying by 100. Results indicated a high degree of accuracy for both paired stimulus formats (80% and 70%) and poor accuracy for the survey (55%). Thus, they demonstrated strong evidence of construct validity—the extent to which preference predicts efficacy—for the two paired stimulus formats by assessing correspondence between preference and reinforcer assessment outcomes for the entire array of items.

Evidence of construct validity decreases the extent to which reinforcer assessments must be conducted. That is, if a preference assessment is consistently found to predict reinforcer efficacy with 100% accuracy, there is little need to conduct a reinforcer assessment following preference assessment administration. This outcome has benefits in applied settings, where time, clinical, or logistical constraints may make it impractical to conduct reinforcer assessments.

Behavioral Economic Assessments of Reinforcer Effectiveness

One factor that is likely responsible for any variance between preference and reinforcer assessment outcomes is differential response effort between the assessments. The increased response effort in the reinforcer assessment, although minimal, may affect relative reinforcer value (DeLeon et al., 2013). That is, the degree to which an individual values one item relative to another may vary as a function of the amount of effort or work required to gain access to the item. DeLeon et al. (1997) examined the effect of increased work requirements on relative preference using items that were similarly preferred on a preference assessment. During the reinforcer assessment, the two items were equally preferred at low work requirements but, as the work requirement for both increased, participants showed a clear preference for one item over the other. This finding and others have demonstrated that reinforcers may be differentially effective under different factors such as work and effort requirements, delay to receipt, and reinforcer quality (e.g., Athens & Vollmer, 2010; Neef et al., 1994). Thus, in addition to examining reinforcer efficacy, it is also important to consider a reinforcer's *effectiveness*[2], or the

[2] The degree to which an item functions as a reinforcer under increasing constraints is commonly referred to as efficacy (e.g., Gilroy et al., 2018; Hursh et al., 2013) and effectiveness (e.g., DeLeon et al., 2013). However, for consistency within this paper, efficacy will refer to a yes–no classification of whether an item functions as a reinforcer, and effectiveness will refer to reinforcing property under varying dimensions of reinforcement.

degree to which it functions as a reinforcer under different levels of these various contributing factors (DeLeon et al., 2013).

Behavioral economics is a discipline that seeks to identify, describe, and measure effects of different variables on relative reinforcer effectiveness (Hursh, 1980, 1984; Lea, 1978). Behavioral economic demand is a methodology for assessing the effectiveness of a commodity across increasing constraints (Reed et al., 2015). Increasing constraints typically refers to increasing price of a commodity—price is often monetary but can also be considered as work or effort (Hursh et al., 2013). Extensive behavioral economic demand research in areas such as treating drug and alcohol abuse has demonstrated that the relative effectiveness of a commodity is susceptible to changes in its price (Bickel et al., 2014). These findings mirror those observed by DeLeon et al. (1997). Specifically, this research has demonstrated that effectiveness is subject to aspects of the environmental context, such as the price of substitutable or complementary reinforcers and the availability of the reinforcer outside of the studied environment (Hursh et al., 2013).

Several studies have applied a behavioral economic framework to examinations of relative reinforcer effectiveness with individuals with intellectual and developmental disabilities (e.g., Call et al., 2012; DeLeon et al., 2009; Frank-Crawford et al., 2018; Peterson et al., 2016; Roane et al., 2005; Tiger et al., 2010; see Gilroy et al., 2018; Roane, 2008). Outcomes from these studies are consistent with behavioral economic research in other domains. For example, Kodak et al. (2007) and Roane et al. (2005) demonstrated that reinforcers that are freely available outside of the therapeutic setting (i.e., an open economy) are less effective when used in a treatment program than reinforcers that are not available outside of the treatment program (i.e., a closed economy), a finding which has been well demonstrated in laboratory settings (Hursh et al., 2013; cf. Timberlake & Peden, 1987). Given the various aspects of the environmental context that affect reinforcer effectiveness across a range of work requirements, Reed et al. (2015) recommend supplementing preference assessments with some form of demand (i.e., effectiveness) assessment.

One method for assessing reinforcer effectiveness is with progressive ratio schedules of reinforcement (Roane, 2008). Unlike a fixed ratio schedule, in which the schedule value (i.e., the number of responses required to earn an item) remains constant, in a progressive ratio schedule, the schedule value increases after each time a reinforcer is earned (Hodos, 1961). Specifically, a progressive ratio schedule places three contingencies on responding: (a) a reinforcer is delivered following completion of the schedule value, (b) the schedule value increases each time a reinforcer is earned, and (c) session time is extended with each response (Jarmolowicz & Lattal, 2010). The primary measure of reinforcer effectiveness from a progressive ratio schedule is break point, which is defined as the last schedule value that was successfully completed (Hodos, 1961).

When used in reinforcer assessments, progressive ratio schedules allow for a rapid assessment of reinforcer effectiveness. Roane et al. (2001) demonstrated the applied utility of progressive ratio schedules in treatment of destructive behavior of four individuals with intellectual and developmental disabilities. Specifically, they found that when used in a treatment package, reinforcers with higher progressive ratio break points produced greater reductions of problem behavior than reinforcers with relatively lower break points. Progressive ratio reinforcer assessments can also be used to assess the degree to which preference assessments measure reinforcer effectiveness (i.e., construct validity), in a similar manner to Northup et al. (1996) for reinforcer efficacy (e.g., Reed et al., 2009). The utility of these findings extends to applied settings in a variety of contexts where it is beneficial to have highly effective reinforcers (e.g., reinforcement schedule thinning, teaching complex skills).

Preference Assessment Methodology in OBM

Although most preference and reinforcer assessment research has been conducted with early education and clinical populations (Virués-Ortega et al., 2014), the methodology and findings from the existing literature can have utility for understanding preference in work-related contexts.

In a popular OBM textbook, Daniels and Bailey (2014) outline an approach for identifying employees' preferred incentives, which is similar to the early stages of a conducting preference assessments for clinical populations (DeLeon et al., 2013). Daniels and Bailey state that the first step is to gather information through collaborative conversations between management and employees to identify a broad array of possible incentives. From that array, managers can select items deemed to be reasonable and sustainable and present them to employees using a formal preference assessment. The preference assessment that they recommend is a survey, wherein incentives are listed, and employees are asked to rate their preference for each incentive on a Likert-type scale from 0 (*none at all*) to 4 (*very much*). Incentives with a higher numeric score are considered an employee's HP incentives; these outcomes can be used to select incentives for individual or group-based incentive programs.

The survey method proposed by Daniels and Bailey (2014) is structurally similar to the format examined by Northup et al. (1996), which was found to have poor accuracy with respect to predictions of reinforcer efficacy with individuals with intellectual and developmental disabilities. One contributing factor for the poor accuracy with that format is likely the items were never present (i.e., items were depicted as pictures) nor delivered at any point during the assessment (Kuhn et al., 2006). However, a survey may be a viable method for assessing incentive preferences for neurotypical adults. Employees responding to the survey would likely have a greater learning history with providing subjective ratings across various formats (e.g., pain scales, movie reviews, online product reviews, customer satisfaction surveys), which may lead to greater accuracy than observed by Northup et al. Unfortunately, at the time when the incentive survey assessment was first proposed (in a previous edition of Daniels & Bailey, 2014), no reinforcer assessment was conducted to demonstrate the survey's accuracy.

To provide a summary of the various types of preference assessments used in OBM, their reliability and validity, and general experimental arrangements, I conducted a systematic review of the literature.

Literature Review Method

I conducted a systematic review of preference assessment methodology in OBM research following the standards for Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA; Moher et al., 2009). Records were identified through searches of PsycINFO, PubMed, and Web of Science through 1/23/20, with no restrictions on publication year. The keywords used for the search included the Boolean operators: "preference assessment*" AND ["employee*" OR "staff"]. The scope of the search was for inclusion of the keywords anywhere in the full text of the article. In addition to the database search, I reviewed the reference list of all records identified by the initial search to identify additional records.

Studies included in the screening had to meet the following criteria: (a) be published in a peer-reviewed journal, (b) full-text available (PubMed database only), and (c) available in English. The study selection procedure is displayed in Figure 1. Initial search results yielded 52, 78, and 15 records from PsycINFO, PubMed, and Web of Science, respectively. An additional seven records were identified through hand searches of reference lists for a total of 152 records.

After removing 57 duplicate records, the 95 remaining records were screened for eligibility based on the following criteria (a) conducted empirical evaluations, (b) involved staff or simulated staff as participants, and (c) performed an assessment of preference for incentives. Initial title and abstract screening identified 84 records that did not meet the search criteria (including six of the seven from the hand search). Of the records removed, one was not in English, seven were review papers and thus not empirical evaluations, 39 did not involve staff or simulated staff, 34 did not include any form of preference assessment, and three included a preference assessment for reasons other than identifying potential incentives (e.g., identifying preferred work locations or activities).

Full-text review of the remaining 11 articles identified one article that did not meet the inclusion criteria (no preference assessment was conducted), leaving 10 articles that met the inclusion criteria. One article (Wine, Reis, & Hantula, 2014) featured two experiments, and another (Wine et al., 2013) conducted two sets of preference assessments, so these articles were classified as containing two studies each. Thus, the review included 12 studies across 10 articles.

The included articles were all published between 2006 to 2014, with nine articles published in the *Journal of Organizational Behavior Management* and one in the *Journal of Applied Behavior Analysis* (Wine & Wilder, 2009). Except for one article (Waldvogel & Dixon, 2008), all studies were conducted by two research groups, Wilder and colleagues and Wine and colleagues.

Articles were rated along the following categories, preference assessment formats, criteria for determining HP and LP incentives, and reinforcer assessment. Those studies that included a reinforcer assessment were also summarized based on experimental design and analysis; participants, setting, and work task; incentives assessed; reinforcement schedule; and outcomes. In addition, I summarized the evidence for test–retest reliability for studies including more than one preference assessment administration, content validity for studies using more than one preference assessment format, and construct validity for studies directly assessing preference assessment outcomes.

Study Characteristics

Preference Assessment Stimuli. Array sizes for the reviewed studies ranged from four to 10 items. Items used in each study are listed in Appendix A. A word cloud depicting relative frequency of each word from the list of incentives used is shown Figure 2. The figure depicts all words that appear three or more times in the list. The words that appear most frequently are "card" (n = 45), "gift" (n = 42), "tickets" (n = 9), "movie" (n = 8), "candy" (n = 8), and "store" (n = 8). The word cloud was created with the R programming language, version 3.6.2 (R Core Team, 2019), using with the tm package for text mining (Feinerer & Honrik, 2019) and wordcloud2 package (Lang & Chien, 2018).

Preference Assessment Formats. Across the 12 studies, four preference assessment formats were used. Five studies used only one format (Wilder et al., 2011; Wilder et al., 2007; Wine & Axelrod, 2014; Wine et al., 2012; Wine & Wilder, 2009), six included two formats (Waldvogel & Dixon, 2008; Wilder et al., 2006; Wine et al., 2013, Studies 1–2; Wine, Kelley, & Wilder, 2014; Wine, Reis, & Hantula, 2014, Experiment 2), and one included three formats (Wine, Reis, & Hantula, 2014, Experiment 1).

Due to inconsistent naming conventions, I classified assessment formats based on each study's description of the assessment methodology rather than the name used by the authors. The format used most frequently was a ranked-choice assessment (hereafter referred to as *ranking*), which was used in nine studies (Waldvogel & Dixon, 2008; Wilder et al., 2011; Wilder et al., 2007; Wine et al., 2012; Wine et al., 2013, Studies 1–2; Wine, Kelley, & Wilder, 2014; Wine, Reis, & Hantula, 2014, Experiments 1–2). The ranking asked participants to rank the array of items from most- to least-preferred, with the highest-preferred item receiving a rating of 1.

The second most frequently used format was a survey assessment (hereafter referred to as *survey*), which was used in eight studies (Wilder et al., 2006; Wine & Axelrod, 2014; Wine et al., 2013, Studies 1–2; Wine, Kelley, & Wilder, 2014; Wine, Reis, & Hantula, 2014, Experiments 1–2; Wine & Wilder, 2009). The survey was based off of the assessment provided by Daniels and Bailey (2014). The survey asked participants to rate each item on a Likert-type rating scale, where higher scores on the rating scale indicate higher preference for that incentive. Seven studies used a five-point rating scale, from 0–4, and one study used a four-point rating scale, from 1–4 (Wine & Wilder, 2009).

The other two formats used methods adapted from those commonly used in therapeutic and educational settings. Two studies used an MSWO (Waldvogel & Dixon, 2008; Wine, Reis, & Hantula, 2014, Experiment 1), where, instead of choosing physical incentives, experimenters wrote the name of each item on an index card and asked participants to select from the array of index cards. Both studies repeated the full MSWO procedure three times and preference hierarchy was determined based on the percentage of times that each item was chosen (i.e., the number of times chosen divided by the number of times presented). The final format used was the paired stimulus assessment, which was used in one study (Wilder et al., 2006). The paired stimulus assessment was adapted for use in this context (i.e., adult participants with strong verbal repertoires). Experimenters listed each possible item-pairing in a random order on a piece of paper and asked participants to circle one item in each pairing. A preference hierarchy was determined based on the percentage of times that each item was chosen.

The specific language used in the assessment varied across and within preference assessment formats. For example, some assessments asked participants to rate items based on how much they prefer each item, whereas others asked how much work they would do to earn each item. Of the nine rankings, five asked participants to list items according to preference, two according to how much work they would do for each item, and two asked for a list from “best” to “worst.” Of the eight surveys, two asked participants to rate items based on preference, four based on how much work they would do for each, and two did not describe the specific language used. Both MSWOs asked participants to choose the item they preferred the most, and the paired-stimulus assessment included the prompt, “which would you do a lot of hard work to get?” The wording used can be classified into four categories: (a) preference, (b) amount of work, (c) best/worst, and (d) unspecified. There appear to be no systematic trends within research group nor across time. Moreover, of the seven studies that included more than one preference assessment, only four used a consistent language style across formats. No studies provide a rationale for the choice of language in the preference assessment, but even subtle changes may produce significant effects on responding (e.g., Henley et al., 2017).

High- and Low-Preference Classifications. Six studies categorized preference assessment outcomes by designating HP and LP incentives. These designations were subjective as no clearly established criteria exist for determining specific preference classifications. As a result, the methods used vary across studies.

Two studies used a standardized criterion across formats. Wilder et al. (2006) defined HP items as those with at least 75% of the maximum rating for the survey (i.e., a rating of 3 or 4 on a

0–4 scale) or items selected on more than 75% of opportunities in the paired stimulus assessment. Wine, Reis, and Hantula (2014, Experiment 2) used a different approach to standardization by conducting a median split of the results from the survey and ranking, such that for each assessment half of the items in the array were classified as HP and half as LP.

The four remaining studies that classified high-preference stimuli used criteria unique to each assessment. Wine, Reis, and Hantula (2014, Experiment 1) was the only study to classify outcomes of the MSWO. For each session, an item's selection percentage was determined by dividing the number of times it was selected by the number of times it was presented and multiplying by 100 (e.g., the item chosen on the third presentation was selected once and presented three times). Then, a mean percentage across all items was calculated and HP items were classified as those items with a selection percentage greater than the mean.

Both studies providing classifications for rankings (Wine, Kelley, & Wilder, 2014; Wine, Reis, & Hantula, 2014, Experiment 1) defined the top three items out of an eight-item array (i.e., the top 75%) as HP incentives. Wine, Kelley, and Wilder (2014) also classified the bottom three items as LP incentives.

Three studies (Wine & Axelrod, 2014; Wine, Kelley, & Wilder, 2014; Wine, Reis, & Hantula, 2014, Experiment 1) used a classification system similar to Wilder et al. (2006) in that items with ratings of 3 or 4 on a 0–4 scale were classified as HP. Likewise, two of those studies (Wine & Axelrod, 2014; Wine, Kelley, & Wilder, 2014) classified items with a rating of 0 or 1 as LP items. Wine and Wilder (2009) used a similar approach that was based on a 1–4 scale. Items with a rating of 1 were LP, ratings of 2 or 3 were medium-preferred, and 4 were HP.

One challenge when establishing criteria for HP and LP designations is that the different preference assessment formats assess preference in different contexts. The survey assesses an *absolute* preference level, where ratings for all items are anchored to specific values (e.g., "none at all" for a value of 0); whereas, the ranking, MSWO, and paired stimulus assessments provide a *relative* order of preference, where items are rated relative to one another. The relative-preference measures are somewhat volatile. The designations may change based on which items are included in the array. For example, an item ranked sixth out of an eight-item array would be considered LP, but if twelve arbitrary (i.e., very LP) items were added into the array, it would be ranked sixth out of 20 and thus would be considered HP.

Nonstandard classifications limit the extent to which comparisons between formats may be made. For example, consider if the classification system for one assessment systematically designated more LP items than the classification system for another assessment. Given research demonstrating that LP incentives often function as reinforcers under low work requirements in single-operant arrangements (e.g., Lee et al., 2010; Penrod et al., 2008; Taravella et al., 2000), reinforcer assessments using such a format would lead to more true-positive and false-negative outcomes (and fewer false-positive and true-negative outcomes). Specifically, items classified as HP are more likely to produce true-positive outcomes than false-positive outcomes, and items classified as LP are more likely to produce false-negative outcomes than true-negative outcomes. As such, the classification system with a bias toward more LP classifications would also have more false-negative outcomes and thus a lower accuracy score than the other system. Thus, when using measures that rely on HP–LP classifications (e.g., accuracy) to compare validity of preference assessment formats, researchers should consider using a standardized classification system across formats.

Reinforcer Assessments. Five studies included reinforcer assessments. Three of the studies used the reinforcer assessments to validate outcomes from multiple preference assessment outcomes (Wilder et al., 2006; Wine, Reis, & Hantula, 2014, Experiments 1–2), and the other two studies used a single preference assessment to identify incentives to be used in evaluations of mixed reinforcer arrangements (Wine & Axelrod, 2014; Wine & Wilder, 2009). All studies included at least the survey as a preference assessment format. In addition, Wilder et al. (2006) also included the paired-stimulus assessment; Wine, Reis, and Hantula (2014, Experiment 1) included the ranking and MSWO; and Wine, Reis, and Hantula (2014, Experiment 2) included the ranking.

The method for selecting items to be included in the preference assessment varied across studies. One study asked participants to nominate items that they would like to earn, and experimenters selected items from that list based on cost and ease of delivery (Wilder et al., 2005). In two studies, experimenters asked employees not participating in the experiment what they would like to receive for completing extra work and selected items based on cost and feasibility (Wine, Reis, & Hantula, 2014, Experiments 1–2). Two studies did not report how they selected items to include in the array (Wine & Axelrod, 2014; Wine & Wilder, 2009). Finally, within each of the five studies, participants completed preference assessments for the same array of items. That is, item arrays were not individualized to participants.

The reinforcer assessment for one study (Wine & Axelrod, 2014) used an ABAB reversal design, where the A phase was a no-reinforcement baseline and the B phase was a reinforcer assessment with only one condition (parametrically assessing different levels of probability of HP incentive delivery). The four remaining studies used multielement designs with initial no-reinforcement baseline phases where the conditions in the multielement represented the opportunity to earn a different incentive. The only study to include all items from the preference assessment array in the reinforcer assessment was Wilder et al. (2006). Wine, Reis, and Hantula

(2014) included only HP items in Experiment 1 and all HP items plus one LP item in Experiment 2. Wine and Wilder (2009) used one HP, one medium-preference, one LP item, and one varied-preference condition for which an HP, medium-preference, or LP item would be randomly selected.

In all reinforcer assessments, participants could earn incentives by completing a given work requirement. The work requirement was individualized to each participant based on their work rate during the initial baseline phase. Across all studies, work requirements remained at a static level throughout the experiment (i.e., all reinforcer assessments used fixed ratio schedules of reinforcement). Thus, these studies are best considered assessments of reinforcer efficacy. All studies set work requirements equal to the mean work units completed per session during baseline—two of the studies calculated mean excluding the final three sessions in baseline, which all had zero-level responding (Wine, Reis, & Hantula, 2014, Experiment 1; Wine & Wilder, 2009). In Wine, Reis, and Hantula (2014, Experiments 1–2), participants could earn only one incentive per session. In all other studies, participants could earn multiple incentives per session.

One study was conducted with simulated staff participants—volunteer research participants from a university—in a laboratory setting resembling an office and a simulated check processing computer program as the work task (Wine & Wilder, 2009). The remaining four studies involved actual employees as participants, with sessions conducted in an office or similar setting located within their workplace, and simple, standardized work tasks that resembled typical tasks that participants would complete in their regular job duties (i.e., filing papers, completing behavioral data sheets). Participants for Wilder et al. (2006) were administrative assistants at a university; participants for Wine and Axelrod (2014) and Wine, Reis, and Hantula (2014, Experiments 1–2) were direct care staff at a human service setting. Although four studies involved employees as participants, their actual workplace behavior was not a focus of the research.

In all four studies, only the target behavior was made available. There were no programmed alternative or off task responses that participants were told they could engage in during sessions. Rather, for three of the studies sessions ended when participants indicated that they did not want to complete additional work (Wilder et al., 2005; Wine, Reis, & Hantula, 2014, Experiment 1; Wine & Wilder, 2009). The fourth study used a similar arrangement, except sessions ended after a predetermined session length (Wine, Reis, & Hantula, 2014, Experiment 2). Although these arrangements are consistent with many monetary-incentive studies, results may be affected by the limited availability of off-task activities, which are often present in natural workplace settings (Oah & Lee, 2011). Thus, it may be beneficial to include off-task activities nonmonetary-incentive reinforcer assessments.

Psychometric Properties

Interrater Reliability. Interrater reliability refers to the degree to which two or more individuals provide similar ratings on an assessment of the same construct (VandenBos, 2015). For a preference assessment, interrater reliability may be considered the extent to which different individuals provide similar ratings for a single individual's preferences. Across two studies, Wilder and colleagues (2007, 2011) assessed the degree to which manager predictions of their employees' preferred incentives corresponded with employees' own rankings (i.e., interrater reliability). They asked 127 employees and 20 managers across varied industries to rank individual employees' preferred incentives. To quantify manager–employee reliability, they calculated Kendall rank-order correlations. Correlations were low in both studies, mean $\tau = .11$

(range, −.6–1), *N* = 27 (Wilder et al., 2007), mean τ = .25 (range, −1–1), *N* = 100 (Wilder et al., 2011). Findings from both studies suggest that managers perform poorly at predicting incentives for employees, which lends support to the need to conduct objective and systematic preference assessments.

Test–Retest Reliability. With respect to preference assessments, test–retest reliability is a measure of the preference stability over time, determined by comparing results from two or more administrations of a single preference assessment format (VandenBos, 2015). Two studies assessed stability of employee preferences over time. Wine et al. (2012) asked 10 direct-care staff to complete rankings for an array of nine incentives once a month for seven months. Kendall rank-order correlation coefficients between each employee's initial ranking and each subsequent monthly ranking were generally low, with a mean τ = .26 (note that data were reported as an *R*, but Kendall is the clearly specified method). However, visual inspection of data provided (Wine et al., 2012, Figure 1) suggests that results may have been idiosyncratic across participants, which is consistent with previous preference-stability research (e.g., Zhou et al., 2001).

In a follow-up study, Wine, Kelley, and Wilder (2014) assessed changes in preference across four different time intervals. Ten participants completed a survey and ranking five times for an array of eight items. The first interval between administrations was four weeks and for each successive interval the duration decreased by one week. Thus, they assessed changes in preference at 1- through 4-week intervals. Analyses of preference stability were based on the number of HP–LP switches—when a participant's rating for an item changed from an HP to LP or LP to HP classification. Results showed no systematic difference across interval durations, except for the 1-week duration, in which neither the survey nor ranking had any HP–LP switches. Across all intervals, Wine et al. observed fewer than one HP–LP switch per participant. These findings provide some support for the test–retest reliability of the survey and ranking assessments to the extent that items classified as HP and LP do not frequently switch designations.

Taken together, data from both studies reflect similar findings from longitudinal preference assessment research conducted in clinical contexts. For example, Zhou et al. (2001) observed relatively low rank-order correlations among 22 adults with intellectual and developmental disabilities, but also noted greater consistency with stimuli designated as HP. These findings suggest that employee preferences are likely to change over time, and thus supervisors would find it advantageous to assess preference frequently, a finding and recommendation which is similar to previous preference assessment research (e.g., Hanley et al., 2006).

Finally, the relative stability of HP and LP incentives lends support to the reliability of the survey and ranking formats. However, this finding should be viewed as preliminary as there were limited sample sizes with no within-subject replication (for the given time intervals), no reinforcer assessments were conducted to validate preference rankings, and only two studies that assessed test–retest reliability.

Criterion Validity. Criterion validity is typically assessed by examining correspondence between an assessment and an established standard (VandenBos, 2015). Given the novelty of research examining incentive preference assessments, there is no established standard with which comparisons can be made and, as such, criterion validity cannot be formally assessed. However, given the varied preference assessment formats in use, it may be beneficial to measure correspondence between them as a form of preliminary criterion validity.

Seven studies included two or more preference assessments. Of those seven studies, only one directly compared correspondence between two or more preference assessment formats (Waldvogel & Dixon, 2008). A second study did not conduct a comparison but included data that could be extracted and analyzed (Wilder et al., 2006). Two further studies did not conduct comparisons but reported information about which incentives were classified as HP. These data allowed for an analysis of agreement between formats (Wine, Reis, & Hantula, 2014, Experiments 1–2). The final three studies were not assessed as they reported neither individual preference assessment outcomes nor comparisons between assessment formats. These studies reported ratings averaged across participants (Wine et al., 2013, Studies 1–2) or coefficients of preference stability (i.e., number of HP–LP classification switches; Wine, Kelley, & Wilder, 2014). In sum, the preference assessment outcomes of four studies were assessed for criterion validity. Specifically, this was considered an assessment of concurrent validity because, within each study, the different preference assessments formats were conducted at the same point in time.

Waldvogel and Dixon (2008) conducted Spearman rank-order correlations between outcomes of the ranking and MSWO for four participants. They found a strong positive correlation between the formats, mean R_s = .8 (range, .60–1.0). Wilder et al. (2006) included the survey and paired-stimulus assessments. I extracted ordinal-rank data from the article (Wilder et al., 2006, Figure 1) and conducted a Kendall rank-order correlation[3] to measure association between the two formats. Across four participants, there was a relatively strong relation between the survey and paired-stimulus preference assessments, mean τ_b = 0.82 (range = .46–.97).

[3] The method and rationale for this test are the same as described in the Method section under Data Analysis.

Correspondence between assessment outcomes used by Wine, Reis, and Hantula (2014, Experiments 1–2) was calculated as item-by-item agreement between each preference assessment pairing for each participant. Agreements were scored if both formats categorized an incentive in the same manner (i.e., HP or not HP), and disagreements were scored if one format categorized an incentive as HP and the other categorized it as not HP. An agreement score was calculated by dividing the number of agreements by the sum of agreements plus disagreements and multiplying by 100. In Experiment 1, mean agreement across participants (n = 3) was 79.2% (range, 50–100) between the survey and ranking and 66.7% (range, 59–87.5) between the survey and MSWO. The manuscript did not provide enough information to calculate agreement between the ranking and MSWO. In Experiment 2, agreement between the survey and ranking was 87.5% for one participant and 100% for the other. Across all five participants in both experiments, mean agreement between the survey and ranking was 85% (range, 50–100).

Analyses of preference assessment outcomes in the four studies demonstrate high levels of correspondence for all comparisons. Although preliminary, these data may be useful in future assessments of criterion validity (e.g., if one assessment format becomes established through demonstrations of high levels of construct validity).

Construct Validity. Given findings that the survey assessment has been demonstrated to be an inaccurate measure of reinforcing efficacy in clinical contexts (Northup et al., 1996, Northup, 2000) and that self-report data may be unreliable under certain conditions (see Critchfield et al., 1998), it is critical to assess construct validity of the survey and similar formats. Specifically, construct validity would be interpreted as the degree to which there is convergence between preference and reinforcer assessment outcomes (i.e., convergent validity; VandenBos, 2015). That is, a preference assessment would have strong convergent validity when incentives classified as HP are found to be efficacious reinforcers and incentives classified as LP are found to be inefficacious reinforcers.

Five studies included reinforcer assessments, which can be examined for evidence of convergent validity. All five studies included at least one HP incentive in the reinforcer assessment. Two studies, which used only the survey assessment, included only one HP incentive (Wine & Axelrod, 2014; Wine & Wilder, 2009). Both studies demonstrated that the HP incentive was an efficacious reinforcer, as participants met the work requirement and earned at least one HP incentive in every session it was made available across both experiments.

The three remaining studies used more than one preference assessment format and included any incentive classified as HP by at least one format in the reinforcer assessment (Wilder et al., 2006; Wine, Reis, & Hantula, 2014, Experiments 1–2). Across all three experiments, participants met the work requirement for nearly all HP incentives. These experiments also predefined the performance criteria necessary for an incentive to be designated as an efficacious reinforcer, and nearly all HP incentives met the criteria necessary for their respective study. The lone exceptions were observed for three of the four participants in Wilder et al. (2006) who all rated the same incentive (help from an assistant on a task) as an HP incentive on both the survey and paired stimulus assessments, but did not work for that incentive during the reinforcer assessment. Thus, it may be that their understanding of the incentive differed from how the privilege was be administered.

Overall, strong convergent validity of HP incentives was demonstrated across five studies for the survey, two for the ranking, and one for the paired stimulus assessment. One important point of consideration is that in studies that included more than one preference assessment format, the different formats did not always classify the same items as HP. For example, in

Wine, Reis, and Hantula (2014, Experiment 2), one participant's survey identified one more HP incentive than the ranking. All incentives with at least one HP designation were demonstrated to be efficacious incentives, which suggests one Type II error (false negative) from the ranking. Thus, it may be important to also consider the convergent validity of LP incentives.

Four studies included at least one LP incentive in the reinforcer assessment. One study did not directly examine the reinforcing efficacy of LP incentives—LP was part of a mixed-reinforcement package with the HP—and could not be assessed for convergent validity of the LP incentive (Wine & Axelrod, 2014). Of the three studies assessed, two studies (Wine, Reis, & Hantula, 2014, Experiment 2; Wine & Wilder, 2009) included only one LP incentive in the reinforcer assessment—for Wine, Reis, and Hantula (2014), the LP incentive received the lowest rating on both the survey and ranking for all participants. Across both studies, LP incentives produced very low levels of performance, with only a single session (out of 14 total sessions) where a participant completed the work requirement. The third study (Wilder et al., 2006) included all LP incentives in the reinforcer assessment. Results varied for the two preference assessment formats. Items classified as LP by the survey produced generally low performance, with only four out of the 12 LP incentives meeting the study's criteria to be classified as efficacious reinforcers; whereas nine out of the 17 incentives classified by the paired stimulus preference assessment as LP were efficacious reinforcers.

Wilder et al. (2006) was the only study to directly assess the reinforcing efficacy of all items included in the preference assessments. An advantage of including all incentives is that it allows for calculation of accuracy as a quantitative measure of convergent validity. Wilder et al. calculated accuracy of both preference assessment formats using the same procedures as Northup et al. (1996). Across all participants, accuracy was 71% for the survey and 50% for the paired-stimulus assessment. These data demonstrate a high degree of convergent validity for the survey and poor convergent validity for the paired-stimulus assessment.

Overall, fewer studies examined LP incentives than HP incentives, and those studies generally examined a relatively smaller range of LP incentives. The three studies all demonstrate strong levels of convergent validity of LP incentives for the survey, with somewhat limited data. Although results based on the ranking demonstrated convergent validity, the data are too limited to make any conclusions. Finally, convergent validity of LP incentives was relatively poor for the paired stimulus assessment.

Although Wilder et al. (2006) defined HP and LP incentives using similar criteria for both assessment formats, the classification system may have been responsible for the paired-stimulus assessment's low accuracy score. Because the survey produces an absolute preference rating, there was no limit on the number of HP incentives; whereas, the paired-stimulus assessment restricted the number of possible HP incentives, given that only a certain number of incentives can meet a given selection-percentage threshold. Thus, the distributions of HP and LP classifications differed across formats. Specifically, half of the incentives were classified as HP by the survey, whereas less than a third were classified as HP by the paired-stimulus assessment. The different distributions raise a problem when making accuracy comparisons given previous reinforcer assessment research, which has consistently demonstrated that low-preference stimuli often function as reinforcers at low work requirements in single operant arrangements, leading to a greater likelihood of false negatives than true negatives (e.g., DeLeon et al., 1997; Roscoe et al., 1999; Tustin, 1994). Thus, the bias toward LP classifications increases the number of false negatives leading to decreased accuracy for the paired-stimulus assessment.

Summary of Reviewed Studies

The twelve studies included in the review varied widely both in aims and methodology. Two studies evaluated accuracy of managerial prediction (Wilder et al., 2007; Wilder et al., 2011), two studies measured preference stability (Wine et al., 2012; Wine, Kelley, & Wilder, 2014), four studies evaluated incentive topographies and reinforcement arrangements (Wine & Axelrod, 2014; Wine et al., 2013, Studies 1–2; Wine & Wilder, 2009), and four studies compared preference assessment formats (Waldvogel & Dixon; Wine et al., 2006; Wine, Reis, & Hantula, 2014, Experiments 1–2).

Overall, there was a high degree of correspondence between all preference assessment formats, and preference assessments provided strong predictions of reinforcer efficacy. However, factors related to the variability in methodology limited the degree to which comparisons could be made and results could be synthesized. These factors included (a) inconsistent HP–LP classifications across studies and between formats and (b) different criteria for defining efficacious reinforcers. These dichotomous rating systems (i.e., dividing a hierarchy of values into yes–no groups) were necessary as the primary means for assessing reinforcer efficacy—in some studies, they were also used to compare preference stability (Wine et al., 2013) and to compare outcomes between formats (Wine, Reis, & Hantula, 2014, Experiments 1–2). However, given the evidence demonstrating that relative reinforcer effectiveness shifts as a function of changes in factors such as delay or work and effort requirements (e.g., DeLeon et al., 1997; Neef, 1994; Tustin, 1994), the utility of the dichotomous rating systems may be limited. That is, it may be more useful to assess how preference assessment outcomes align with reinforcer effectiveness across increasing work requirements, thereby removing the necessity of defining reinforcer value based on arbitrary criteria. Assessing reinforcer effectiveness may also have greater practical utility; from an organizational standpoint, supervisors are likely primarily concerned with how preference assessment outcomes align with the amount of work that employees would be willing to do to earn an incentive.

Although research has demonstrated correspondence between preference assessment outcomes and measures of relative reinforcer effectiveness among clinical populations (e.g., DeLeon et al., 2009), to date no research has examined this relation in an organizational context. However, recent research has demonstrated an application of behavioral economic demand to work performance. Henley et al. (2016) evaluated the utility of a hypothetical work task (HWT) for assessing employees' willingness to work for a given incentive. The HWT is a questionnaire that asks participants whether they would be willing to complete a given number of work units in exchange for a monetary incentive ($0.05 or $0.10). Participants provide a yes-or-no response, and if they respond "yes," the question is re-presented with a higher number of work units; this process is continued until participants respond "no." The highest number of work units that a participant is willing to complete is considered their break point. Thus, the HWT is similar to a progressive ratio reinforcer assessment, except responses are self-reported rather than observed and the incentive is not actually delivered. Results from Henley et al. demonstrated that participant's demand for the incentive decreased in a manner consistent with existing behavioral economic research (e.g., Hursh et al., 2013).

The HWT is an adaptation of a hypothetical purchase task, which is an assessment method used to assess behavioral economic demand in a variety of commodities (e.g., alcohol, opiates, food; Roma et al., 2017). Hypothetical purchase task methodology was developed as a means of assessing the value of drug reinforcers without having to deliver the actual consequence (Jacobs & Bickel, 1999; Murphy & MacKillop, 2006; Petry & Bickel, 1998).

Behavioral economic researchers have demonstrated a high degree of correspondence between self-reported hypothetical-purchase-task responses and observed behavior (Amlung & MacKillop, 2015; Roma et al., 2015). In addition, Henley (2017) demonstrated a high degree of correspondence between break points on an HWT and break points on a progressive ratio reinforcer assessment—during which participants performed the task and received real monetary incentives. These initial findings suggest that the HWT may provide a useful measure of reinforcer effectiveness. Although Henley et al. (2016) assessed only one incentive (i.e., money), the HWT may have utility as a preference assessment in work-related contexts. An HWT could be completed for each item in an array, which would provide a hierarchy of break points directly related to the targeted performance. Taken together, these findings and gaps in the incentive-preference-assessment literature with respect to reliability and validity of the preference assessments suggest that it would be worthwhile to examine the extent to which the HWT and existing incentive preference assessment formats provide accurate measures of reinforcer efficacy and effectiveness.

Purpose

The purpose of the present experiment is to assess (a) correspondence between the survey, ranking, and HWT preference assessment formats; (b) correspondence between measures of preference and measures of reinforcer efficacy and reinforcer effectiveness; and (c) preference stability over time.

Method

I compared outcomes of three staff-level preference assessment methods: a survey, a ranking, and an HWT. Next, I examined the construct validity of the preference assessments using two reinforcer assessment methods: a fixed ratio assessment and a progressive ratio assessment. Finally, I conducted follow-up preference assessments to evaluate preference stability over the course of the experiment.

Participants

Six female undergraduate students enrolled in a large Midwestern university participated in the experiment. They were recruited from applied behavior analysis classes via in-class announcements (Appendix B). The university's human research protection program approved all methods prior to recruitment (#00141691). Compensation for participation was in the form of course credit. Participants earned extra credit toward their final average, and there was a maximum amount of extra credit available in the course so that students unable to participate in this study could still earn up to the maximum amount through several other methods (e.g., participating in survey or in-person research experiments, additional assignments). Participants also had the opportunity to earn the incentives that were built into the experiment. Table 1 lists demographic and compensation information for all participants. Study procedures took place over four to six 1-hr blocks, and average total duration across participants was 4.0 hr (range, 3.5–4.8).

Setting

Data collection occurred in a small research room located on the university's campus. The room measured 2.2 by 2.0 m and was arranged to simulate an office space or cubicle (see Appendix C for a diagram of the room layout). The primary workspace included a Dell computer monitor (48.1 cm diagonal), keyboard, mouse, and two small speakers set up on a folding table. The research room was adjoined to an observation room by a one-way mirror. The computer monitor, keyboard, mouse, and speakers were connected to a Dell desktop computer located in the observation room. A mirrored monitor, keyboard, and mouse, which were connected to the same desktop computer, were located in the observation room to allow the experimenter to remotely set up the experimental conditions.

Materials and Preference Assessment Stimuli

A variety of items were located throughout the session room to simulate an office environment (e.g., wall-mounted maps of the United States and the world, a clock, and phone chargers). Several leisure items were placed on the table, to the right side of the computer monitor. These items included puzzle books (e.g., crossword, sudoku), playing cards, a fidget spinner, and colored pencils. In addition, participants were allowed to bring any outside items into the session room (e.g., phone, laptop, homework, food).

Preference assessment stimuli were chosen by first identifying potential items through informal discussion with undergraduate students about items that they would be willing to work for. Next, items from those discussions were identified that were feasible for the experimental context, similar in topography to those used in incentive preference assessment studies, and with an approximate value of $1–$3. The items were “chocolate,” “fruity candy,” “gel pen,” “gift card,” “gum,” and “lip balm.” Four of the items (chocolate, fruity candy, gum, and lip balm) allowed for participants to choose between two related options. For example, participants who earned chocolate would have the opportunity to choose between either a Twix or Snickers bar. These choice options were included with the goal of maximizing the probability that at least some incentives were valuable to each participant. The specific items included are described in Appendix D. During the reinforcer assessments, I delivered incentives using 10.5-by-10.5-cm laminated paper tokens that displayed an image and text of the items with which they were associated (displayed in Appendix E). Tokens were exchanged for backup reinforcers at the end of each 1-hr block. Tokens allowed for immediate reinforcer delivery following each session while minimizing potential session-to-session changes in responding from satiation that could occur if participants were able to consume the incentives immediately.

Apparatus

All experimental procedures were delivered through the computer using a program written with Microsoft Visual Basic. The computer was a Dell Vostro 260S running Windows 8.1. The monitor was a Dell 1909W with a display resolution of 1440 by 900 pixels (landscape orientation) and a 60 Hz refresh rate. I developed the program as a Windows form application using the .NET Framework 4.6.1. The program occupied the entire screen when run; participants could neither close the program nor access other programs on the computer. The program contained two broad application types: an experimental work task and the preference assessment delivery.

Work Task

The work task was an abstract task, with a single work unit comprised of three component tasks. Broadly, for each work unit, the program presented a target code that would be used through the three component tasks where participants were to: (1) type the target code in a box, (2) click on the box in an array of three that is paired with the target code, and (3) click on a second matching box that is moving across the screen. The objective in creating the work task was to create a discrete work unit with fixed duration and difficulty across multiple presentations of the units that was also more complex than those commonly used in OBM research (e.g., check processing, paper sorting) to facilitate greater sensitivity to differences in reinforcer value (Madden & Perone, 1999). In addition, the task was designed to be arbitrary, such that there would be no inherent or perceived value in completing work. The arbitrary nature of the task was intended to (a) control for participant history and (b) minimize the amount of reinforcement outside of the programmed incentives that participants might contact for completing the task.

Specifications

At the start of each new work unit, the program randomly generated three unique four-character alphanumeric codes. All codes consisted of two letters and two integers selected from a range of A–Z and 1–9, respectively. The order of letters and numbers within the code was unconstrained. Each code was randomly paired with one of three colors: *blue*, *red*, and *yellow*. Finally, one code–color dyad was randomly selected as the *target stimulus* for the work unit, with the other code–color dyads designated as *distractor stimuli*.

To facilitate discrimination between alphanumeric characters, codes and numbers (i.e., points, time remaining) were displayed in a monospace typeface and letters used in the codes were displayed in uppercase—the monospace typeface also ensured a standard width for the code display. The program automatically converted all entered letters to uppercase, so letters entered without the Shift or Caps Lock keys were displayed and interpreted as uppercase letters. The standard computer shortcut keys Enter (to submit) and Tab (to place the cursor inside a text-entry box) were disabled, thereby requiring use of the mouse for those actions (other keys and keymaps that remained active included Backspace, Delete, Home, End, and Control + left/right arrows).

Participant errors (described below) produced a blackout screen for 1.5 s. The blackout screen was a semitransparent (97% opacity) black screen with white text stating, “Please wait.” Participants could not interact with the program while the blackout screen was present. The purpose of the blackout screen was to reduce the possibility that participants could complete the work unit without attending to the relevant stimuli (e.g., rapidly clicking the mouse and cycling through all stimuli). That is, the blackout screen placed a putative negative punishment contingency on errors.

Work Task Completion

The program began by presenting an instruction screen with text stating, “For the next 3-minute session, you will have the opportunity to complete the work task. You may do as much or as little work as you would like, including no work at all. Press ‘START’ to begin the session and then feel free to do whatever you want for the remainder of the session.” When participants pressed the start button, the session began, and the program displayed the work task for the remainder of the session.

Figure 3 displays the interface presented to participants as they progressed through each component of a work unit. The first component of the work unit involved typing the target code into a text-entry box. Instructions positioned at the top left of the screen stated, “Enter the code into the box below.” The target code was positioned at the bottom-right of the screen, and the text-entry box was positioned below the instructions at the top-left of the screen.

To complete the first component, participants had to use the mouse to click inside the text-entry box, use the keyboard to type the code, and then use the mouse to click on the “submit” button located directly below the text-entry box. If the code was entered incorrectly when participants pressed submit, the program recorded an error and displayed the blackout screen for 1.5 s. After the blackout, the program cleared the text-entry box, and participants could start the component over. If the code was entered correctly, the program advanced to the second component. For the remainder of the work unit, the first-component instructions were no longer visible, and the text-entry box and button were disabled but remained visible (i.e., participants could not interact with these elements).

The second component involved selecting the target color (i.e., the color paired with the target code) out of an array of three stationary code–color dyads. Instructions positioned at the bottom-left of the screen stated, “Click on the color above that matches the code.” Positioned directly above the instructions were the three code–color pairings. The pairings were arranged in three rows, with codes in the left column and 1-by-1-cm (35-by-35-pixel)[4] colored squares in the right column. The order of the colors—and thereby the position of the target-stimulus dyad—was randomly determined on each work unit.

To complete the second component, participants had to use the mouse to click on the target-color square. If participants clicked on the incorrect square or clicked anywhere on the screen other than the target square, the program recorded an error and displayed a blackout screen for 1.5 s. After the blackout, participants could attempt the second component again. If they clicked on the target-color square, the program advanced to the third component. For the remainder of the work unit, the second-component instructions were no longer visible.

The third component involved using the mouse to select the target-color square as it moved across the screen. Instructions positioned at the bottom-left of the screen stated, “Click on the moving box on the right that corresponds with the color and code.” Positioned in the center of the screen was a 20-by-20-cm (700-by-700-pixel) black field. The field contained three squares of the same size and colors as those displayed in the second component, and each square independently moved within the black field. The squares moved diagonally across the field at a speed of 6.1, 7.1, or 8.1 cm/s (212, 250, or 283 pixels/s). The speed and angle at which squares moved was determined by randomly assigning each square an independent vertical and

[4] In addition to descriptions of size as stimuli appeared for participants, all physical dimensions of program elements are indicated in pixels for a more precise unit of measurement (35 pixels are displayed on the screen as approximately 1 cm).

horizontal speed of either 4.3 or 5.7 cm/s (150 or 200 pixels/s). When a square reached the top or bottom of the field, it was randomly assigned a new vertical speed in the opposite direction; when a square reached the left or right side of the field, it was randomly assigned a new horizontal speed in the opposite direction. For example, if a square reached the top of the field, its vertical direction would change from up to down, and its vertical speed would be randomly assigned as 4.3 or 5.7 cm/s (150 or 200 pixels/s). This process ensured that two squares did not travel in the same or a similar path. Movement of the squares was not affected by position of other squares, so it was possible for squares to overlap one another. When a target-color square overlapped with a distractor-color square, the target-color square was always positioned in front of the distractor-color square.

To complete the third component, participants had to use the mouse to click on the target-color square moving within the black field. If they clicked anywhere on the screen other than the target-color square, the program recorded an error and displayed a blackout screen for 1.5 s. After the blackout, participants could attempt the third component again (the squares continued to move throughout the blackout). If they clicked on the target-color square, the program played a soft chime sound and added one point to their score for that session.

After completing the third component, the program started a new work unit with a new target code. The program continued presenting the work task for the entire duration of a given session. At the end of a session, the program displayed a black feedback page with white text. The specific wording of the feedback varied with the experimental conditions, but all variations included the following statements (a) “The session has ended”; (b) a summary of points scored, number of incentives earned, both, or neither; and (c) “Please turn off the monitor and wait for the experimenter.”

Response Measurement and Definitions

For all preference assessments, I recorded assessment-specific ratings for each incentive. For the survey, higher number ratings were considered higher preference incentives; for the ranking, lower number ratings were considered high preference; and for the HWT, the break points (i.e., highest work amount that participants indicated they would complete) for each incentive were recorded, with higher break points corresponding to a higher preference.

I defined each participant's HP and LP incentives using a similar approach as Wine, Reis, and Hantula (2014), taking a median split of participant ratings on each assessment. This method of classifying incentives standardized the classifications across the three preference assessment formats, so that each format would yield an equal proportion of HP incentives. The three top-rated items were classified as HP incentives, and the three bottom-rated items were classified as LP incentives. In the case of ties, all tied items were classified as HP. I also recorded assessment duration for all preference assessments to assess whether there were any meaningful differences in implementation time.

The primary dependent variable was *points* scored during each session. One point was recorded when participants completed all three components of the work unit. Data were automatically collected via the computer for all components of the experiment.

Analyses of reinforcer assessment outcomes included measures that were derived from points scored. For the fixed ratio reinforcer assessment, I assessed reinforcer efficacy by calculating the percentage of sessions in which a participant completed the work requirement for each incentive. Efficacious reinforcers were classified as incentives that maintained performance above the work requirement for at least two-thirds of the sessions in which they were presented in the reinforcer phase of the fixed ratio assessment; all other incentives were classified as inefficacious reinforcers for the purposes of calculating preference assessment accuracy. For the progressive ratio reinforcer assessment, a break point was defined as the highest schedule value successfully completed during a progressive ratio session (Hodos, 1961). Finally, cumulative points were the cumulative sum of points scored for all sessions of a given condition within a phase.

General Procedure

The experiment took place over five stages: (1) orientation, (2) preference assessments, (3) fixed ratio reinforcer assessment, (4) progressive ratio reinforcer assessment, and (5) follow-up preference assessments. Participants completed all of the orientation, the preference assessments, and the first several sessions of the fixed ratio assessment in the first 1-hr block. The second 1-hr block consisted entirely of sessions in the reinforcer assessment phase of the fixed ratio assessment. The third 1-hr block typically consisted of the final sessions in the fixed ratio assessment and the first several sessions of the progressive ratio assessment. The final 1–2 blocks consisted of the remainder of the progressive ratio assessment and the follow-up preference assessment. After the follow-up preference assessments, the experimenter conducted an unstructured debrief in which participants were asked why they worked or did not work. Notes taken from debrief statements are provided in Appendix F. Any tokens earned during a block were traded in for their respective incentives at the end of each hour block. All session blocks occurred on separate days.

Orientation Procedure

Study procedures began with participants completing an informed consent form (Appendix G) and brief demographic questionnaire (Appendix H). Next, the experimenter reviewed participants' schedule and compensation.

To familiarize participants with the work task, they started by completing two 3-min practice work-task sessions. First, the experimenter explained that, during parts of the study, participants would have the opportunity to complete a simulated work task on the computer. The experimenter prompted participants to read the instructions on the computer screen, press “Start,” and follow the prompts on the screen, which described how to complete the task. Participants then completed the first session, during which time the experimenter was not present in the room. Following the first session, the experimenter asked the participant if they had any questions, then instructed them to complete the second session using the time to build fluency with the task.

Next, to mitigate potential demand characteristics—which have been reported as a potential confound in incentive research (Oah & Lee, 2011)—participants completed two 3-min off-task sessions. The experimenter explained that the participant would have the option to work or not work and clarified that at any point during the experiment they could engage with any of the off-task stimuli present in the room, use their phone, laptop, or do anything else (participants could not use the computer for anything other than the program). Participants then completed the two off-task sessions, where they were instructed to press “Start” to begin the work task, but spend the entire session doing any activity other than work. That is, they were instructed to start the work task and then not complete any work for the session. The purpose of these sessions was to provide a forced-choice exposure to off-task behavior in the context of the experimental arrangement.

Preference Assessment Procedure

I compared outcomes of three preference assessments: a survey assessment, a ranking assessment, and an HWT. The survey and ranking used wording that is identical to that used by

Wine, Reis, and Hantula (2014), and the HWT followed the format used by Henley et al. (2016), with adaptions for use as a preference assessment. All three preference assessments were delivered by the computer program. The assessments were delivered in the same order for all participants, with the survey first, then the ranking, then the HWT. Following each preference assessment, participants completed a brief math distractor task. Figure 4 displays the interface displayed for the preference assessments.

Preference assessments were administered twice during the experiment. Initial preference assessments took place following the orientation stage, and follow-up preference assessments took place at the end of the experiment, following the progressive ratio reinforcer assessment. The follow-up preference assessments were presented in the same manner as the initial preference assessments, with the exception that the incentive stimuli were not re-introduced during follow-up.

The preference assessment stage began with the experimenter showing participants the array of preference assessment items and defining the names used to describe each of the items (e.g., chocolate means a choice between a Snicker or Twix bar). The experimenter also explained that when a participant earned an incentive, they would be given a token for each incentive earned, and the tokens would be exchanged at the end of the daily lab visit.

The experimenter prompted participants to read the initial instructions and complete the questionnaires that would be presented and then left the room while participants completed the preference assessments. Prior to presenting the first preference assessment, the program displayed the following instructions:

> In the following sections you will be asked how much work you would do to earn the various incentives. For these questions, work refers to the number of points you would be

willing to score in the sample task. Please answer honestly, thoughtfully, and as if you were actually in the situation.

After reading the instructions, participants pressed an "advance" button and the program presented the first preference assessment.

Survey

On the survey, a textual prompt positioned at the top of the screen asked participants, "How much work would you be willing to do to earn each item?" The six incentives were listed in alphabetical order in a column positioned below the instructions. To the right of each incentive were the numbers 0–4 with ordinal anchors listed above the numbers stating *None at all*, *A Fair Amount*, and *Very Much* for the numbers 0, 2, and 4, respectively.

To complete the survey, participants had to use the mouse to provide a rating for all six incentives and click the advance button positioned on the lower-right of the screen. If they did not provide a rating for all incentives, the program prompted them to do so. Once participants pressed advance, the program asked them to confirm that they wanted to submit their responses, then advanced to the first math distractor task.

Distractor Task

The math distractor task was included to mitigate potential carryover effects from one preference assessment format to the next by providing an intervening behavior for participants to engage in. It is similar to distractor tasks used in episodic memory research (e.g., Aue et al., 2017) and was used with the goal of weakening any stimulus–stimulus pairings formed between incentives and numeric ratings on previous assessments. However, the direct effects of the task were not evaluated, so this is a putative function. The task prompted participants to calculate a running summation of a series of numbers. The initial instructions stated, "Add the following numbers to the best of your ability." These instructions were presented for 10 s along with a progress bar to visually depict time remaining until the task started. The program then presented a sequence of 20 single-digit, positive integers, each present on the screen for 3 s. I used three number sequences—one for each of the three distractor-task presentations—which were randomly generated to include the numbers 1–9, with the constraint that no number could be presented consecutively. Sums for the first, second, and third distractor task were 114, 103, and 107, respectively. After all numbers in the sequence were presented, the program prompted participants to enter the sum into a text-entry box and press submit. Thus, the task duration was 70 s plus the time taken to enter the sum and press submit. Once participants pressed submit, the program advanced to the next preference assessment.

Ranking

For the ranking, instructions positioned at the top of the screen stated, "Please rank the following items in order based on how much work you would do to earn each item." Directly below the instructions were labels indicating, "1 = I would do the most work to earn this item" and "6 = I would do the least work to earn this item."

Participants could indicate their rankings using six blank dropdown boxes, numbered 1–6, which were arranged in an ascending order in a column below the instructions. Clicking on a dropdown box with the mouse revealed a list of the six incentives in alphabetical order, and participants could make a selection by clicking one of the incentives in the list. Once an incentive was selected in a dropdown box, it no longer appeared in the lists for the other boxes; this prevented the possibility of listing the same incentive at multiple rankings. On the right side of each dropdown box was a "reset" button, which, when pressed, cleared the dropdown box, allowing participants to rank the incentive in a different location. Finally, a "reset all" button cleared all dropdown boxes, allowing participants to restart the ranking procedure. Once participants pressed advance, the program asked them to confirm their responses and advanced to the second math distractor task. Following the second math distractor task, the program advanced to the HWT.

Hypothetical Work Task

Participants completed an HWT for each of the six items, presented sequentially in alphabetical order. Instructions positioned at the top of the screen stated, “For each question below, please indicate whether you would score the number of points indicated to earn the specified item.” Additional text below the instructions asked, “Would you be willing to score 1 point in exchange for one [incentive name]?” Below the question were two buttons: one labeled “yes” and one labeled “no.” The buttons were 0.9 cm (32 pixels) in height, 2.3 cm (81 pixels) in width, and were positioned side-by-side, separated by 0.5 cm (16 pixels), with the yes button to the left of the no button. When the HWT was first displayed, the program positioned the mouse cursor in between the two buttons. After each yes–no response, the program repositioned the cursor back to the original starting position, such that, for each response on the HWT, the cursor always began equidistant from the two buttons.

If participants chose yes, the number of points displayed in the question increased to the next value in the HWT progression. If participants chose no or reached the end of the HWT progression, the program cleared the screen and recorded the break point for that incentive (i.e., the last value that participants indicated they would work for). After a brief 2 s delay, the program restarted the HWT for the next incentive starting back at the first HWT value. The program repeated this procedure until all incentives were presented.

The progression of ratio values in the HWT was: 1, 3, 5, 8, 11, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60. These values were selected using the following parameters: (a) the number of ratio values (i.e., 15) is the same as used previously by Henley et al. (2016); (b) the first value is equal to 1; (c) the median ratio value (i.e., 25) was set to an approximate average point total in a series of 3-min pilot sessions; and (d) step sizes were no greater than 5 to facilitate a greater degree of break-point sensitivity. After all incentives were presented, the program advanced to the third and final math distractor task. Following the third distractor task, the program advanced to a blank screen indicating that the session was over.

Fixed Ratio Reinforcer Assessment

Procedure

The experimenter began the fixed ratio reinforcer assessment by instructing participants that during the next portion of the experiment, they may feel free to do as much or little work on the work task as they want, including none at all. The experimenter then read the presession instructions that were displayed on the computer aloud and prompted participants to begin the first session.

Design

I used a multielement design with an initial baseline phase (Johnston & Pennypacker, 2009) to compare the reinforcing efficacy of all six incentives. All sessions were 3 min in duration, during which time participants had the opportunity to complete the work task to earn a designated incentive. For all participants, I conducted a minimum of three baseline sessions and continued the phase until I observed stable responding based on visual inspection of the data. In the reinforcer assessment phase, I conducted a minimum of three sessions for each incentive condition.

Baseline. During the baseline phase, there was no incentive for completing the work task. The purpose of this phase was to establish a baseline work rate, which would be used to set the work requirement in the reinforcer assessment phase. Prior to the start of the session, the program displayed instructions stating, "For the next 3-minute session, you will have the opportunity to complete the work task. You may do as much or as little work as you would like, including no work at all. The incentive for the session is listed below." The program then stated the incentive as "None" and the minimum number of points to earn the incentive as "N/A." Once participants pressed the start button, the program started the session and displayed the work task.

At the end of the session, the program displayed a summary page stating, "The session has ended. You scored [point value] point(s). There was no incentive for this session. Please turn off the monitor and wait for the experimenter." The experimenter then entered the room and stated that for the previous session there was no incentive, so they did not earn anything; the experimenter then prompted the participant to continue with the next session.

Reinforcer Assessment. During the reinforcer assessment phase, participants had the opportunity to earn one incentive per session. The purpose of this phase was to assess the reinforcing efficacy of each incentive. In addition to the six incentives, I included a baseline condition, which was identical to sessions in the baseline phase. Thus, there were a total of seven incentive conditions (six incentives plus baseline). For each session in this phase, the program presented one of the seven condition types in a counterbalanced, quasi-random order. In addition, no condition was presented two consecutive times, and one session for each of the six incentives was presented prior to the first baseline-condition session in this phase; except for this constraint, baseline sessions were in a random position in the session order for the phase. Session orders are listed in Appendix I.

Prior to the start of the session, the program displayed instructions that were identical to those in the baseline phase with the exceptions that (a) the incentive label indicated which incentive would be available for that session and (b) the minimum number of points needed to earn the incentive label indicated participants' respective work requirements. Note that for baseline-condition sessions, the instructions were identical to those in the baseline phase.

The work requirement, or the number of points necessary to earn an incentive, was individually determined for each participant and remained at a fixed level throughout the fixed ratio assessment. It was set to equal one point more than the highest single-session point total in the baseline phase. This method differed from that used in previous incentive-preference-assessment research (e.g., Wine, Reis, & Hantula, 2014), where work requirements were equal to the mean level of responding during baseline. I set the work requirement based on the maximum level of responding to be more consistent with the functional definition of a reinforcer as a stimulus that, when delivered contingent on a behavior, increases the rates of that behavior.

Five participants did not score any points during baseline phase sessions. For these participants, I set the work requirement equal to half (rounded up to the nearest whole number) of the points scored in the orientation session with the highest point total. I based this criterion off of the orientation sessions because a work requirement of one point would not produce a meaningful assessment of reinforcer efficacy. However, during orientation sessions, participants had been prompted to work the entire time; thus, using the highest point total from orientation may have been too high a work requirement.

Session duration remained at 3-min for all reinforcer assessment phase sessions. If participants completed the work requirement before the end of the session, the program still continued to present new work units until the end of the 3-min session. There were no additional contingencies for completing more work than the work requirement. Participants could only earn one incentive per session.

At the end of the session, the program displayed a summary page stating, "The session has ended. You scored [point total] point(s)." If participants completed the work requirement, the program stated, "You earned 1 incentive" and provided a prompt to turn off the screen. The experimenter then entered the room, restated the number of points earned and that the participant earned the incentive, and delivered one token for the respective incentive. If participants did not meet the work requirement, the program stated, "You did not earn the incentive." The experimenter feedback similarly restated the point total and that the participant did not earn the incentive. For baseline-condition sessions, the summary page and experimenter feedback were identical to the baseline phase.

Progressive Ratio Reinforcer Assessment

The progressive ratio assessment was structured in a similar manner to the fixed ratio assessment, with the primary modification to how reinforcers could be earned. In this assessment, participants could earn incentives by completing the work task along a progressive ratio schedule of reinforcement.

Procedure

The experimenter began the progressive ratio reinforcer assessment by instructing participants that the next portion of the experiment would be different from the previous stage in that they would have the opportunity to earn multiple incentives over a 10-min session. The experimenter then read the presession instructions that were displayed on the computer aloud and prompted participants to begin the first session.

Progressive Ratio Schedule Progression

Each participant had a unique progressive ratio schedule made up of selected values from the HWT. Table 2 lists progressive ratio schedules for all participants. I selected values for each participant based on their responding during the fixed ratio assessment using the following parameters: (a) it would be possible to complete three to five ratio values by working continuously throughout a single session; (b) the first value was set at approximately midpoint between the work requirement and the lowest observed levels of responding in the fixed ratio assessment; and (c) the second ratio value was the first HWT value greater than the work requirement in the fixed ratio assessment. For example, during the fixed ratio assessment, Julianne’s work requirement was 10 and her lowest observed level of responding was 0, so the first value in her progressive ratio progression was 5 and her second value was 11. Note that because Shannon’s responding during the fixed ratio assessment was greater than the work requirement, I omitted the second parameter and used the third parameter to set the first value. These parameters allowed for (a) multiple opportunities to earn an incentive in a single session, which creates the possibility for differential work rates across sessions; (b) an assessment of incentive value at lower work requirements than in the fixed ratio assessment; and (c) an assessment of values much greater than the initial fixed ratio work requirement.

Design

I used the same design for the progressive ratio assessment as with the fixed ratio assessment, a multielement design with an initial baseline phase (Johnston & Pennypacker, 2009). Session duration varied as a function of participant responding in each session. Sessions continued until 1 min elapsed with no completed work units, with a maximum session duration capped at 10 min. For all participants, I conducted three sessions in the baseline phase and three sessions of each incentive condition in the reinforcer assessment phase.

Baseline. During the baseline phase, no incentive was provided for completing the work task. The purpose of this phase was to serve as a comparison to work rates during the reinforcer assessment phase. In addition, baseline was included to expose participants to the progression of the progressive ratio schedule prior to the first incentive session. Instructions displayed prior to the start of the session stated, "For the next session, you will have up to 10 minutes and can earn multiple incentives. You may do as much or little work as you would like, including no work at all. If you do not want to do any more work, you can just stop, and the session will end after a brief period of time." The program then stated the incentive as "None" and prompted participants to begin the session by pressing the start button. Once they pressed the start button, the program began the session and displayed the work task.

The work task interface was modified to reflect the progressive ratio contingencies. Instead of indicating how many points participants had scored in the session, the program displayed the number of points they needed to score to complete the schedule value. The program also displayed the number of incentives that participants had earned during a session. Although there was no incentive during baseline conditions, the program still displayed points needed and incentives earned as if there were an incentive.

At the end of the session, the program displayed a summary page stating, "The session has ended. There was no incentive for this session. Please turn off the monitor and wait for the experimenter." The experimenter then entered the room and stated that for the previous session there was no incentive, so they did not earn anything; the experimenter then prompted the participant to continue with the next session.

Reinforcer Assessment. During the reinforcer assessment phase, participants had repeated opportunities to earn a designated incentive within each 10-min session. For each session, the program presented one of the seven condition types (six incentives plus baseline) in a counterbalanced, quasi-random order, following the same constraints as in the fixed ratio assessment. Session orders are listed in Appendix J.

Instructions prior to and at the end of the sessions were identical to those presented during baseline with the exception that the program stated "You earned [number of incentives earned] incentive(s)" during the incentive conditions. At the end of the session, the experimenter entered the room, stated how many incentives participants earned, and delivered one token for each incentive earned.

Data Analysis

To supplement visual inspection, I conducted several analyses to assess reliability and validity of the preference assessments. All statistical analyses were conducted using the R programming language, version 3.6.2 (R Core Team, 2019). In addition, I used the Kendall package (McLeod, 2011) for Kendall correlation analyses and the sjmisc package (Lüdecke, 2018) to conduct median splits for the accuracy calculations.

I calculated Kendall rank-order correlation coefficients to measure ordinal association between all preference assessment and reinforcer assessment outcomes. Specifically, due to the presence of tied values in the data, I calculated the tau-b coefficient (Siegal, 1956, Equation 9.10). All correlation significance tests were two-tailed. Like other correlation methods (e.g., Pearson, Spearman), Kendall coefficients range from −1–1, with a correlation of 1 indicating a perfect relation between two variables. However, there are no specific guidelines for interpreting

Kendall correlation coefficients. Rather, coefficients may be interpreted by calculating statistical significance of the coefficient.

The outcome used for the fixed ratio assessment was the proportion of sessions that participants completed the work requirement for each incentive. The outcome used for the progressive ratio assessment was average break point for each incentive, which was calculated by summing the break points for all sessions of a given incentive and dividing by three. These correlation coefficients were used to assess (a) correspondence between all three preference assessment formats; (b) the convergent validity of the preference assessments, or the extent to which the preference assessment outcomes corresponded with fixed and progressive ratio assessment outcomes; and (c) the test–retest reliability of each of the preference assessments over time.

I also calculated the accuracy of HP–LP classifications from each preference assessment based on fixed ratio assessment outcomes using the same method as Northup et al. (1996) and Wilder et al. (2006). True positives were defined as instances where efficacious reinforcers were classified as HP; false negatives were instances where efficacious reinforcers were classified as LP. False positives were defined as instances where inefficacious reinforcers were classified as HP; true negatives were instances where inefficacious reinforcers were classified as LP.

Results

Individual Participants

Julianne

Julianne’s data for the initial and follow-up preference assessments and fixed and progressive ratio reinforcer assessments are depicted in Figure 5. Her initial preference assessment data revealed a high degree of correspondence between formats, with the primary discrepancy occurring with the gift card. For all three formats, she reported that she would do the most work for lip balm and gel pen. She also rated the gift card as an item she would complete a relatively high amount of work for on the survey and HWT but ranked it second lowest on the ranking. On all three formats, she reported that she would do the least work for gum. A median split classified lip balm, gel pen, and gift card as the HP items from the survey and HWT. The ranking assessment also classified lip balm and gel pen as HP items, but included fruity candy in place of the gift card.

For the fixed ratio reinforcer assessment, Julianne did not complete any work during the baseline phase, so her work requirement for the reinforcer assessment phase was set to 10 points based on her performance during orientation. In the reinforcer assessment phase, she did not work during any sessions in the baseline condition. For the two incentives rated as LP on all three formats (gum and chocolate), she completed the work requirement in just one of six combined sessions. For the two incentives rated as HP on all three formats (lip balm and gift card), she completed the work requirement for in all six combined sessions. Notably, during sessions when Julianne completed the work task, she always earned exactly the number of points needed to complete the work requirement and then stopped working for the remainder of the session.

During the progressive ratio assessment, Julianne did not complete any work during the baseline phase. In the reinforcer assessment phase, she did not complete any work for any sessions in the baseline, chocolate, fruity candy, and gum conditions. Chocolate and gum were rated at LP on all three preference assessments; fruity candy was rated as LP on two assessments. She completed some work during the gel pen and lip balm conditions, earning one to two incentives in each session with those items, which were the two items rated as HP on all three assessments. She completed the most work for gift card, which was rated as HP by the survey and HWT. Julianne's average break point for gift card sessions was 18.3 points, which was more than double the next highest average break point, 8.3 points for lip balm.

For the follow-up preference assessments, there was excellent correspondence between Julianne's ratings for the survey, ranking, and HWT. For all three formats, she reported that she would do the most work for gift card and the least work for gum. A median split classified gift card, lip balm, and gel pen as HP items on all three assessments.

Shannon

Shannon's data for the initial and follow-up preference assessments and fixed and progressive ratio reinforcer assessments are depicted in Figure 6. Her initial preference assessment data revealed excellent correspondence between the three preference assessment formats. Excluding ties, all three formats produced identical ratings. In all formats, she reported that she would do the most work for gum, gift card, and fruity candy and the least work for chocolate. A median split classified gum, gift card, and fruity candy as her HP incentives.

During the fixed ratio reinforcer assessment, Shannon did not complete any work during the baseline phase, so work requirement for the reinforcer assessment phase was set to 11 points based on her performance during orientation. In the reinforcer assessment phase, she did not complete any work for any sessions in the baseline condition. For all other incentive conditions, she completed the work task and earned the incentive, with little differentiation in level of responding among incentive types.

For the progressive ratio assessment, Shannon did not complete any work during the baseline phase nor during the baseline condition of the reinforcer assessment phase. In the reinforcer assessment phase, she completed some work for all incentives, earning one to three incentives per session. The item that she completed the least work for was chocolate, which received the lowest rating on all three preference assessments, with an average break point of 16.7 points. She completed the most work for the gift card, which was highly rated on all three formats, with an average break point of 30 points. Average break point for all other incentives ranged from 21.7–26.7 points.

Due to a procedural error that occurred prior to her sixth session in the reinforcer assessment phase, she completed one session with an incorrect progressive ratio schedule (the schedule progression was 5, 11, 15, 20, 30, 35). This error occurred for a session in the chocolate condition, and she earned four incentives with a break point of 20 points. This session was not included in any data analyses, and the chocolate session with the correct schedule was conducted immediately after. However, the reinforcement contingency was still upheld, so she received tokens for that session. Thus, her responding during chocolate conditions in the progressive ratio assessment may have been affected by the additional chocolate that she earned due to this error.

For the follow-up preference assessments, there was excellent correspondence between Shannon's ratings for the survey, ranking, and HWT. For all three formats, she reported that she would do the most work for the gift card and the least work for chocolate and fruity candy. A median split classified gift card, lip balm, and gel pen as HP items from the ranking and HWT; the survey also included gum as an HP item.

Alexia

Alexia's data for the initial and follow-up preference assessments and fixed and progressive ratio reinforcer assessments are depicted in Figure 7. Her initial preference assessment data revealed excellent correspondence between the three preference assessment formats, with nearly identical relative ratings for each. For all three formats, she reported that she would do the most work for the fruity candy, lip balm, and chocolate and the least work for gel pen and gift card. A median split classified fruity candy, chocolate, and lip balm as her HP incentives.

For the fixed ratio reinforcer assessment, Alexia did not complete any work during the baseline phase, so her work requirement for the reinforcer assessment phase was set to 10 points based off of her performance during orientation. In the reinforcer assessment phase, she did not complete any work for any sessions in the baseline condition. For two incentives rated as LP on all three formats (gel pen and gum), she completed the work requirement in just one of six combined sessions. However, she completed the work requirement for all sessions for the gift card, which was also rated as LP on all three formats. She completed the work requirement for all sessions with the three HP incentives (Fruity candy, lip balm, and chocolate). During sessions that she completed work, she completed only the amount necessary to earn the incentive or slightly higher.

During the progressive ratio assessment, Alexia completed enough work to complete the first schedule value in the first session and then did not work for the remainder of the baseline phase. During the reinforcer assessment phase, she did not complete any work for any sessions in the baseline or gel pen conditions. Gel pen was her lowest rated item on all preference assessments. She completed the most work for the gift card, which was her second lowest rated item on all three preference assessments, with an average break point of 18.3 points. Average break point for all other incentives ranged from 8.7–13.7 points.

For the follow-up preference assessments, there was excellent correspondence between Alexia's ratings for the survey, ranking, and HWT. For all three formats, she reported that she would do the most work for the gift card and the least work for the gel pen. A median split classified gift card, fruity candy, and chocolate as HP items from the survey and ranking; the HWT also included lip balm and gum as HP items.

Cindy

Cindy's data for the initial and follow-up preference assessments and fixed and progressive ratio reinforcer assessments are depicted in Figure 8. Her initial preference assessment data demonstrated a high degree of correspondence between formats. On the survey, she reported that she would do the most work for chocolate, the gift card, and lip balm. On the ranking assessment, she indicated she would do the most work for chocolate, whereas, on the HWT, the gift card was the item with the highest break point. The item she indicated she would do the least work for on all three assessment formats was gel pen. The primary discrepancy in her data is that her break point for chocolate was low relative to how she rated it on the survey and ranking—note that chocolate was the first incentive presented on the HWT, which may have affected the low break point. A median split classified chocolate, gift card, and lip balm as the HP items from the survey and ranking; the HWT included the same items in addition to gum.

For the fixed ratio reinforcer assessment, Cindy did not complete any work during the baseline phase. Her work requirement for the reinforcer assessment phase was set to 12 points based on her performance during orientation. In the reinforcer assessment phase, she did not complete any work for any sessions in the baseline condition. For all other incentives, she completed the work task and met the work requirement to earn the incentive for all sessions with the exception of her final gel pen session. Gel pen was rated as LP on all three assessment formats. During sessions that she completed work, she completed only the amount necessary to earn the incentive, with just two exceptions—the first chocolate and last gum sessions, for which she completed approximately double the work requirement.

For the progressive ratio assessment, Cindy worked during the third session in the baseline phase, completing only the initial schedule value. During the reinforcer assessment phase, she did not complete any work for sessions in the baseline condition but worked for all incentive conditions, earning one to four incentives per session. The item she completed the least work for was gel pen, which received the lowest rating on all three preference assessments, with an average break point of 6.7 points. She completed the most work for the gift card, which was among the highest rated on all three preference assessments, with an average break point of 30 points. Average break point for all other incentives ranged from 11.7–15 points.

For the follow-up preference assessments, there was strong correspondence between Cindy's ratings for the survey, ranking, and HWT. For all three formats, she reported that she would do the most work for the gift card and the least work the gel pen. The primary discrepancy was that gum was rated as relatively higher preferred than chocolate on the survey and HWT but rated as less preferred than chocolate on the ranking. A median split classified gift card and fruity candy as HP items on all three assessments. Gum was classified as an HP item by the survey and HWT; chocolate was identified as an HP item by the ranking and HWT.

Elise

Elise's data for the initial and follow-up preference assessments and fixed and progressive ratio reinforcer assessments are depicted in Figure 9. Her initial preference assessment data revealed wide variability across formats. For the survey and HWT, she reported she would do the most work for lip balm and the least work for gift card. However, the ranking assessment revealed the opposite results, where she reported she would do the most work for the gift card and the least for lip balm. A median split classified chocolate, fruity candy, gum and lip balm as the HP items from the survey and HWT. The ranking assessment classified fruity candy, gel pen, and gift card as HP items.

One notable observation is that Elise may have provided her rankings in reverse order, with her lowest preferred item ranked at 1 and her highest at 6. This may have been likely given that the ranking assessment was presented following the survey, in which highly preferred items are given a higher number. During an informal debrief following completion of her participation in the study (six days after the initial ranking assessment), she reported that she did not believe that she could have mixed up the order of the ranking assessment. However, given the response patterns observed from other participants, it seems likely that this may have been the case. Thus, I reordered her ranking assessment data by reversing the rank orders. Elise's data including the reordered rankings are depicted in Figure 10 (note that with the exception of the initial ranking assessment, all data in this figure are identical to those depicted in Figure 9). The median split using reordered ranking data classified chocolate, gum, and lip balm as her HP items—these items were the same as those identified by the survey and HWT.

For the fixed ratio reinforcer assessment, Elise did not complete any work during the baseline phase. Her work requirement for the reinforcer assessment phase was set to 9 points based on of her performance during orientation. In the reinforcer assessment phase, she did not complete any work for the first two baseline sessions, scored 14 points for the third session, and returned to zero points for the fourth session. For the two incentives rated as LP on all three formats (gel pen and gift card), she completed the work requirement in just one of eight combined sessions. For the four remaining incentives, she completed the work requirement in 15 of 16 combined sessions. For all sessions that she completed the work requirement, she completed the work task at levels just above the work requirement, with little differentiation among incentive types.

For the progressive ratio assessment, Elise worked during the second session of the baseline phase, completing only the first schedule value. In the reinforcer assessment phase, she did not complete any work for gel pen or baseline conditions. Gel pen was one of two incentives rated as LP on all three preference assessments. The other incentive rated as LP on all three formats was gift card, which was the only incentive for which she worked for all three sessions in the reinforcer assessment phase, with an average break point of 12.3. For all other incentives, she worked for only the first session in the reinforcer assessment phase.

For the follow-up preference assessments, there was excellent correspondence between Elise's ratings for the survey, ranking, and HWT. For all three formats, she reported that she would do the most work for the gift card and the least work for the gel pen and lip balm. A median split classified gift card, chocolate, and fruity candy as HP items from all three assessment formats.

Marie

Marie's data for the initial and follow-up preference assessments and fixed and progressive ratio reinforcer assessments are depicted in Figure 11. Her initial preference assessment data revealed strong correspondence across preference assessment formats. For all three formats, she reported that she would do the most work for chocolate, fruity candy, and gum and the least work for gel pen. A median split classified chocolate, fruity candy, and gum as her HP items from the ranking assessment. The survey and HWT included the same items in addition to the gift card.

Marie worked during all three sessions of the baseline phase of the fixed ratio assessment, with a maximum session total of 19 points, so her work requirement for the reinforcer assessment phase was set to 20 points. In the reinforcer assessment phase, she worked for the first two of three sessions in the baseline condition and all gel pen sessions but did not meet the work requirement for any session in either condition. Gel pen was rated as LP on all three preference assessment formats. For all other incentives, Marie completed the work requirement at levels just above the work requirement, with little differentiation in level of responding among incentive types.

For the progressive ratio assessment, Marie worked during the first and third sessions of the baseline phase, completing one schedule value in the first session and three schedule values in the third session. During the reinforcer assessment phase, she did not complete any work for baseline condition sessions and worked for only the first of three gel pen sessions. Gel pen, which was her lowest rated item on all three preference assessments, had the lowest average break point of 5 points. She worked the most for chocolate, fruity candy, and gift card, with an average break point of 30 points; these items were all rated as HP by on at least two preference assessments. The remaining incentives, gum and lip balm, had average break points of 28.3 and 21.7, respectively.

For the follow-up preference assessments, there was excellent correspondence between Marie's ratings for the survey, ranking, and HWT. For all three formats, she reported that she would do the most work for the gift card and chocolate and the least work for the gel pen. A median split classified gift card, chocolate, and gum as HP items for all three assessment formats.

Aggregate Analyses

Primary outcome data from the eight assessments (i.e., survey, ranking, HWT, fixed ratio reinforcer assessment, progressive ratio reinforcer assessment, follow-up survey, follow-up ranking, and follow-up HWT) are depicted for each participant in Figure 12–Figure 17. These figures provide a visual depiction of relative incentive value across the different assessment formats. Note that HWT and progressive ratio axis scales are not standardized across assessments. Table 3 displays average Kendall tau correlation coefficients across all participants. Individual participant coefficients are listed in Table 4–Table 9. Data summaries including Elise's original ranking values are depicted in Figure 18 and Table 10.

Preference Assessments (Initial & Follow-up)

Initial preference assessment data for all six participants are depicted together in Appendix K. Overall, I observed a high degree of correspondence between the preference assessment formats for all participants. Kendall tau correlation coefficients measuring the association between the survey and ranking ranged from τ_b = .65–.97, with a mean of τ_b = .83, and statistically significant positive correlations for five of the six participants at the $p < .05$ level. There was also a high degree of correspondence between the HWT and the survey and ranking formats. The mean correlation between the HWT and survey was τ_b = .91 (range, .79–1.0), with statistically significant relations for four participants. The mean correlation between the HWT and ranking was τ_b = .84 (range, .65–.97), with statistically significant relations for four participants. Note that mean correlations between the ranking and other formats were calculated using Elise's reordered data; coefficients for the original data (Table 10) were the same level but in the opposite direction (i.e., demonstrated a negative association). The positive association between the survey and ranking formats is consistent with and adds to the limited research comparing these methods (Wine, Reis, & Hantula, 2014). In addition, the relation between the HWT and the survey and ranking provide some initial evidence in support of its use as a preference assessment.

Follow-up preference assessment data for all six participants are depicted together in Appendix L. The mean correlation between the follow-up survey and follow-up ranking was τ_b = .88 (range, .79–.97), with statistically significant results for five participants. The mean correlation between the follow-up HWT and follow-up survey was τ_b = .89 (range, .70–.96), with statistically significant results for four participants. The mean correlation between the follow-up HWT and follow-up ranking was τ_b = .90 (range, .78–.97), with statistically significant results for five participants. Overall, these strong relations between preference assessment formats replicate those observed in the initial preference assessments, providing further evidence toward demonstrations of concurrent validity.

Although the correlations between preference assessment methods remained high, there was some discrepancy between initial and follow-up ratings. The mean correlation between initial and follow-up surveys was τ_b = .33 (range, −.36–.85), with statistically significant results for Julianne and Marie. The mean correlation was τ_b = .29 (range, −.20–.47) for the ranking and τ_b = .31 (range, −.17–.77) for the HWT, with no statistically significant results for either. Overall, test–retest reliability was generally low for all participants across assessment formats.

Finally, the time required to complete the preference assessments was short for all formats. Average durations across all participants' initial and follow-up administrations was 27.1 s for the survey, 29.7 s for the ranking, and 63.6 s for the HWT (see Table 11 for a detailed breakdown). These durations are consistent those times reported in previous incentive preference assessment studies. Wilder et al. (2005) reported a mean duration of 68 s for the survey, and

Wine, Reis, and Hantula (2014; Experiment 1) reported mean durations of 32 s and 34 s for the survey and ranking assessments (no other studies measured exact durations).

Fixed Ratio Reinforcer Assessment

All six participants' fixed ratio reinforcer assessment data arranged together are depicted in Appendix M. Across all participants, I observed a general pattern where participants either completed no work at all or just enough to complete the work requirement. Once participants completed the requirement to earn the incentive, they stopped working for the remainder of the session. The only participant with any sessions where they completed work but did not meet the requirement was Marie. For those sessions, she briefly engaged in off-task behavior at the beginning of the session and then worked for the remainder of the session. Notably, Marie was also the only participant whose work requirement was determined based on her performance during baseline. Although there were some instances where participants worked well above the requirement (e.g., Cindy, Elise), these instances were limited to no more than one session per condition type.

One measure of convergent validity is the relation between preference assessment ratings and the proportion of sessions that participants earned each item during the fixed ratio reinforcer assessment. The mean Kendall correlation coefficient between the survey and fixed ratio assessment was τ_b = .65 (range, .44–.92), with a statistically significant result for one out of five participants (Julianne). Note that Shannon's data are not included in these analyses because she met the work requirement on every session for all incentives and thus a correlation could not be calculated. The mean correlation between the ranking and fixed ratio assessment was τ_b = .51 (range, .30–.60), with no statistically significant relations for any participants. Finally, the mean correlation between the HWT and the fixed ratio assessment was τ_b = .62 (range, .44–.92), with a statistically significant result for one out of five participants (Julianne).

Correlation coefficients were also high for comparisons with the follow-up preference assessments. The mean correlation for the relation between the fixed ratio assessment and follow-up preference assessments were τ_b = .62 (range, .09–.96) for the survey, τ_b = .63 (range, .30–.93) for the ranking, and τ_b = .59 (range, .24–.92) for the HWT. There were statistically significant relations between the fixed ratio assessment and all three follow-up assessments for Julianne only.

One potential reason why there were so few results with statistically significant relations—despite relatively strong Kendall correlation coefficients—is the limited number of possible values with the outcome variable (i.e., proportion of sessions earning the incentive). For most participants, the possible values were 0, .33, .67, and 1. This limited range increased the likelihood of tied values, which decreases the already-limited power of these analyses. For example, Cindy and Marie had only two unique values, with five of the six incentives tied at 1.0, and Alexia had only three unique values, with four incentives tied at 1.0. Thus, the extent to which convergent validity can be assessed by the correlation analyses may be limited.

The second method for assessing convergent validity is by calculating each preference assessment's accuracy in predicting efficacious reinforcers. Accuracy data are presented in Table 12. Overall, accuracy was high for all three preference assessment formats, with mean values of 77.8% (range, 50.0–100) for the survey, 66.7% (range, 50.0–83.3) for the reordered ranking, and 80.6% (range, 50.0–100) for the HWT. The accuracy rating for the survey was similar to the 71% accuracy that Wilder et al. (2006) obtained with the survey. For all three preference assessments, true positive was the result with the most likely outcome, and false positive was the result with the least likely outcome. Specifically, there was only one instance of a false positive (Julianne's ranking).

Accuracy data for the follow-up preference assessments are presented in Table 13. Overall, accuracy was high for all three preference assessment formats, with minimal differentiation between formats. Mean accuracy scores for the follow-up assessments were 69.4% (range, 50.0–100) for the survey, 66.7% (range, 50.0–100) for the ranking, and 69.4% (range, 50–100) for the HWT. These accuracy data were similar to those observed from the initial preference assessments. Further, for all three preference assessments, true positive was the most frequent outcome and false positive was the least frequent outcome, which is the same pattern observed with the initial preference assessments.

Progressive Ratio Reinforcer Assessment

All six participants' fixed ratio reinforcer assessment data arranged together are depicted in Appendix N, and participants' average break points and cumulative response totals for each incentive are depicted in Table 14.

Kendall correlation coefficients measured the extent to which preference assessment outcomes corresponded with a measure of reinforcer effectiveness, mean progressive ratio break point. The mean correlations for the progressive ratio assessment were $\tau_b = .29$ (range, −.26–.72) for the survey and $\tau_b = .29$ (range, −.30–.75) for the ranking, with no significant relations. The mean correlation between the progressive ratio assessment and the HWT was $\tau_b = .40$ (range, −.26–.85), with a statistically significant result for Marie. Overall, there was a weak positive relation between the preference assessment and reinforcer assessment outcomes.

Correlations were much stronger for the follow-up assessments. The mean correlation between the follow-up survey and progressive ratio break points was $\tau_b = .84$ (range, .67–.96), with statistically significant relations for three participants. The mean correlation between the follow-up ranking and progressive ratio assessment break point was τ_b = .86 (range, .69–.97), with statistically significant relations for four participants. Finally, the mean correlation between the follow-up HWT and the progressive ratio assessment break point was τ_b = .85 (range, .77–.93), with statistically significant results for three participants.

Discussion

This experiment was an initial examination of a laboratory model for assessing incentive preference assessments. I examined the psychometric properties of three incentive preference assessment formats: a survey, ranking, and HWT. Specifically, for each preference assessment format, I assessed criterion validity, construct validity for reinforcer efficacy, construct validity for reinforcer effectiveness, and test–retest reliability.

Correspondence between the survey, ranking, and HWT was strong for all participants across both initial and follow-up preference assessments. These measures of criterion validity were equally strong for all comparisons, across initial and follow-up administrations. These findings are similar to those observed by Waldvogel and Dixon (2008), who compared a ranking and MSWO, and Wilder et al. (2006), who compared the survey and paired stimulus assessment. Overall, the three assessment formats all produced similar ratings of preference.

Despite similarly high levels of correspondence for both initial and follow-up preference assessments, measures of test–retest reliability were poor for all three preference assessments, suggesting a shift in preference over the experiment. These findings reflect outcomes of research with incentive preference assessments (Wine et al., 2012) as well as those in educational and clinical populations (e.g., Hanley et al., 2006; Zhou et al., 2011).

The preference assessments' construct validity with respect to reinforcer efficacy was measured by comparing the ratings with fixed ratio reinforcer assessment outcomes. These measures were high for all three preference assessment formats across initial and follow-up administrations. Coefficients for the ranking assessment were slightly lower than the survey and HWT from the initial administration, but there were no differences between formats in the follow-up administration. Findings were consistent with previous studies that assessed the survey, ranking, or both (Wilder et al., 2006; Wine, Reis & Hantula, 2014) and demonstrate that all three preference assessments provide good measures of the degree to which incentives will function as reinforcers under low work requirements.

The preference assessments' construct validity with respect to reinforcer effectiveness was measured by comparing ratings with progressive ratio reinforcer assessment outcomes. All three preference assessment formats demonstrated similar levels of validity. However, there was a difference in validity levels between initial and follow-up preference assessment administrations. The initial preference assessments for all three formats provided weak measures of validity for reinforcer effectiveness, whereas, the follow-up preference assessments provided strong measures of validity. The discrepancy between initial and follow-up preference assessments is a reflection of the test–retest data, which indicated a shift in preference ratings. Based on the follow-up preference assessment data, the preference assessments provided excellent measures of reinforcer effectiveness. This experiment was the first to examine reinforcer effectiveness of nonmonetary incentives and findings are similar to studies conducted with clinical populations (e.g., DeLeon et al., 2009).

Contributions to the Literature

This experiment contributes to the incentive preference assessment literature in several areas. The primary contribution is the incorporation of a behavioral economic framework into nonmonetary incentive research. This experiment is the first to examine the relation between incentive preference assessment outcomes and reinforcer effectiveness, which is important for translation from laboratory settings to practical application. Although there is utility in identifying which items meet the functional definition of a reinforcer in controlled settings, supervisors planning to implement an incentive program are likely looking to identify which incentives an employee is willing to work the most for. The behavioral economic framework provides a direct measure of reinforcer effectiveness, which can be quickly identified using a progressive ratio reinforcer assessment. This study mirrors similar advances in clinical contexts, which have led directly to improved client outcomes (DeRosa & Roane, 2015).

The incorporation of behavioral economics also addresses a potential constraint in how fixed ratio reinforcer assessment results can be interpreted. In the fixed ratio assessment in the present experiment, participants typically either completed just enough work to earn an incentive or did not work at all, which is similar to previous incentive reinforcer assessments (Wine, Reis, & Hantula, 2014). This response pattern often makes it difficult to identify differentiation in responding between incentives. Undifferentiated results, with several items providing an identical measure, are likely a contributing factor when assessing correspondence in preference assessment results (Kang et al., 2011). Thus, data from the fixed ratio assessment often do not provide levels of differentiation necessary to validate preference assessment results. This limitation is apparent in the finding from the present experiment where average correlations with the fixed ratio assessment were generally high, but few yielded statistically significant results.

By assessing multiple work requirements, the progressive ratio assessment leads to greater levels of differentiation which can assist in validating preference assessment outcomes.

The experiment also provides an initial investigation of the HWT as a potential preference assessment format. As a verbal analog of a progressive ratio reinforcer assessment, the HWT could *potentially* provide a stronger measure of reinforcer effectiveness than the survey and ranking. Although in the present experiment, validity measures were similar for all three assessment formats, future research could provide additional insight and modifications to improve validity. Another potential benefit of the HWT is that its outcome measure (break point) is based on an actual work unit, compared to the arbitrary scale on the survey and relative value from the ranking. The HWT scale allows for a wide range of response outcomes that may provide a more accurate measure of absolute value than a survey. For example, on the survey, Cindy rated all incentives as 0, 1, or 2, whereas, on the HWT she rated her top-preferred item with a break point that was more than double the next highest preferred (Figure 8). The HWT showed a clearer difference in magnitude of preference—as well as a clearer preference hierarchy—than the survey. However, this experiment provides only an initial demonstration of the HWT and more research is necessary before application in practice.

A second contribution is that the experiment directly assessed several psychometric properties of the incentive preference assessment formats. In my review of the extant incentive preference assessment literature, I found few direct examinations of reliability and validity, which were spread out over several articles, with some variability in methodology between studies. This experiment is the first to assess multiple psychometric properties of incentive preference assessments, including comparisons of these measures between different formats. Overall, the present experiment provides (a) the second direct examinations between any pairing of the three preference assessment formats (Waldvogel & Dixon, 2008), (b) the first examination of construct validity of reinforcer effectiveness, (c) additional data that are consistent with measures of construct validity for reinforcer efficacy (Wilder et al., 2006; Wine, Reis, & Hantula, 2014, Experiments 1–2) and test–retest reliability (Wine et al., 2012). Examination of the psychometric properties is valuable for incentive preference assessment research—particularly given the relative novelty of this area of research—as it allows for greater confidence in results when it is not possible to conduct reinforcer assessments and helps shape interpretation of assessment results for both research and practice.

This experiment was only the second to assess the entire array of items in a reinforcer assessment (or a sample of items throughout the hierarchy), which is critical in assessing construct validity. Including items from throughout the preference hierarchy allows for examination of differentiated outcomes between HP and LP items. If only HP items are tested, and all are found to be efficacious reinforcers there is still a possibility that all items in the array would have functioned as reinforcers. For the test to have demonstrated construct validity, it must demonstrate sensitivity to provide discriminative results between efficacious and inefficacious reinforcers. Thus, testing all items in the array gives the clearest assessment of sensitivity of the preference assessments. The inclusion of LP items may also be beneficial when considering practical applications. There are likely many occasions when it is useful to identify a large array of effective incentives, not just a single incentive. Assessing LP items in preference assessment research can aid in understanding the degree to which they function as reinforcers in work contexts.

The final contribution is that this experiment broadly extends on methods for incentive research. It is the first nonmonetary incentive to incorporate alternative responses into the session. Likewise, the methods facilitated sensitivity to reinforcement contingencies in an efficient manner, which may be beneficial for monetary and nonmonetary incentive studies. One OBM study addressed this same concern but used 8-hour workdays over extended visits to accomplish the goal (Oah & Lee, 2011). The present experiment demonstrated that such lengths may not always be necessary.

Limitations and Future Research

The contributions of this experiment notwithstanding, interpretations of results should be tempered with consideration of several limitations and areas that may call for future research. There are four broad considerations with respect to the following areas (a) satiation and incentive classes, (b) preference assessment framing, (c) progressive ratio schedules, and (d) the translational nature of the experiment.

The first limitation is that the experimental arrangement may have facilitated satiation for some incentives. Satiation has been demonstrated as a key contributing factor in changes in preference among individuals with intellectual and developmental disabilities (Hanley et al., 2006) and may have contributed to the variance between preference and reinforcer assessment outcomes. For example, Marie indicated during the debrief that she stopped working for gum toward the end of the experiment because she felt she already earned enough and didn't need any more (Appendix F).

Although satiation can be expected in applied contexts, the translational nature of the study, involving short sessions and frequent opportunities to earn incentives, may accelerate satiation beyond what might occur naturally. Any effect of satiation on preference hierarchy in the present study is likely amplified due to the different classes of incentives used, as some incentives may have been more resistant to satiation than others (Catania, 2013). For example, the gift card, as a generalized conditioned reinforcer, was likely less susceptible to satiation than edible incentives, such as chocolate. Thus, aspects of the experimental arrangement combined with different incentive classes may explain why the gift card emerged as the most effective reinforcer in the progressive ratio assessment, despite being rated as highest-preferred by only one participant (Cindy).

Future research can address issues with satiation through modifications such as using a single class of incentives, using smaller magnitude incentives, or reducing the number of opportunities to earn incentives. In addition, future research examining the differential effects of various incentive classes could provide insight for application. Given that incentive preference may change regularly, it may be difficult for supervisors to regularly assess preference and make new incentives available; thus, it may be more feasible for supervisors to use generalized conditioned reinforcers. Future research can assess reinforcer effectiveness and stability of generalized conditioned reinforcers relative to other classes.

A second limitation related to changes in preference is that the preference assessment language may not have adequately explained the experimental context. Although satiation likely affected relative preference for incentives, it is possible that these changes in preference could have been mitigated with a clearer explanation of the experimental context. That is, if participants were aware of the arrangement in which they would be earning incentives at the start of the experiment, they may have considered the scalability of the gift card in the beginning. This possibility is supported by behavioral economic research involving hypothetical purchase tasks. In a systematic review of a subset of hypothetical purchase task literature (specifically, assessing behavioral economic demand for alcohol), Kaplan et al. (2018) identified several studies demonstrating that participants' responses on hypothetical purchase tasks are sensitive to changes in the language used (e.g., Gentile et al., 2012; Skidmore & Murphy, 2011). Specifically, language used can involve descriptions of context, type and size of reinforcer to be consumed, time period within which it must be consumed, and monetary budget constraints. Given this sensitivity, Kaplan et al. recommend tailoring language specifically to the experimental framework.

Future incentive preference assessment language should include specific information detailing the general contingencies, number of opportunities to earn each incentive, and expectations with respect to availability of alternative activities. Though inclusion of this information is especially useful in contrived, laboratory settings, there is also likely a benefit in practice. Future research can examine whether providing additional detail about how incentives will be implemented improves preference assessment accuracy and stability.

The third consideration is with respect to the implications of using behavioral economics for the study of nonmonetary incentives. An important benefit of using a progressive ratio schedule in reinforcer assessments is that the schedule progression allows for a rapid assessment of reinforcer effectiveness (Roane, 2008). However, in this experiment, the schedule progressions were determined based on responding during the fixed ratio reinforcer assessment. Thus, the process of assessing reinforcer effectiveness in this experiment was relatively time consuming given that the methodology for the progressive ratio assessment relied on completion of the fixed ratio assessment. There may be a possibility that if the progressive ratio schedule progressions were set a priori, participant responding may not be sensitive to a particular schedule progression, which could result in undifferentiated patterns of responding between incentives, thereby limiting any utility of the progressive ratio schedule. Future research can address this limitation by conducting a reinforcer assessment with a progressive ratio schedule that is determined a priori. Results showing differentiated patterns of responding can demonstrate the practicality and usefulness of progressive ratio schedules in nonmonetary incentive research.

Finally, given that this was a laboratory-based, translational experiment, there are several factors limiting the practical application of these findings. Importantly, the work task was abstract and unlike the work that employees in a natural environment would be completing. The task was designed specifically for this experiment with the goal of maximizing sensitivity to changes in reinforcement contingencies. However, in a natural environment, there may be uncontrolled reinforcers associated with work completion (Poling et al., 2000). These uncontrolled reinforcers likely decrease sensitivity to programmed incentives, which would affect how reinforcer assessment results may be interpreted. For example, for an employee who regularly contacts natural reinforcers for their work (e.g., positive client interactions and outcomes), an incentive shown to be moderately effective in a laboratory setting may produce no difference over baseline in the natural setting. In this example, only the most effective incentives would produce meaningful improvements in performance. Thus, future research should extend the methods from this experiment using a more naturalistic work task. It may be especially beneficial to identify a work task that is viable in a laboratory setting that also has a clear translation to the workplace.

An additional consideration is that employees in a natural environment have multiple concurrent job responsibilities, each with their own associated contingencies. Although participants in the present experiment were explicitly informed that they could engage in off-task behaviors, there was only a single work task for which incentives were programmed. As a result, the experiment more closely resembled a single-operant arrangement than the concurrent-operant arrangement of the workplace. Research examining reinforcer effectiveness in single- and concurrent-operant arrangements has consistently demonstrated that effectiveness can be significantly affected by the availability of alternative forms of reinforcement (e.g., DeLeon et al., 2007; Roscoe et al., 1999). Future research should examine the predictive utility of incentive preference assessments for use in concurrent arrangements.

In sum, despite a large body of OBM literature on most-effective monetary incentive arrangements (e.g., Abernathy, 1996; Bateman & Ludwig, 2003; Bucklin & Dickinson, 2001; Dixon & Belisle, 2018), relatively little guidance on nonmonetary incentive assessment and delivery exists. Nonmonetary incentives introduce unique considerations that may not be present for monetary incentives and thus warrant their own line of research. Considerations may include methods for selecting group incentives or the utility of different incentive classes. These and other considerations present a rich opportunity for OBM research, which can draw upon methods from the base of preference assessment literature in educational and clinical contexts addressing functionally similar concerns (DeRosa & Roane, 2015; DeLeon et al., 2013; Piazza et al., 2011).

Important prerequisites for future research examining preference and reinforcer arrangements include a sound methodology for incentive preference and reinforcer assessments and an understanding of the relation between incentive preference assessment outcomes and measures of reinforcer effectiveness. Though translational in nature, the present study helps provide the groundwork toward these goals, which may ultimately help extend the reach and effectiveness of OBM interventions.

References

Abernathy, W. B. (1996). *The sin of wages*. PerfSys Press.

Amlung, M., & MacKillop, J. (2015). Further evidence of close correspondence for alcohol demand decision making for hypothetical and incentivized rewards. *Behavioural Processes*, *113*, 187–191. https://doi.org/10.1016/j.beproc.2015.02.012

Athens, E. S., & Vollmer, T. R. (2010). An investigation of differential reinforcement of alternative behavior without extinction. *Journal of Applied Behavior Analysis*, *43*(4), 569–589. https://doi.org/10.1901/jaba.2010.43-569

Aue, W. R., Criss, A. H., & Novak, M. D. (2017). Evaluating mechanisms of proactive facilitation in cued recall. *Journal of Memory and Language*, *94*, 103–118. https://doi.org/10.1016/j.jml.2016.10.004

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*(1), 91–97. https://doi.org/10.1901/jaba.1968.1-91

Bateman, M. J., & Ludwig, T. D. (2003). Managing distribution quality through an adapted incentive program with tiered goals and feedback. *Journal of Organizational Behavior Management*, *23*(1), 33–55. https://doi.org/10.1300/J075v23n01_03

Bickel, W. K., Johnson, M. W., Koffarnus, M. N., MacKillop, J., & Murphy, J. G. (2014). The behavioral economics of substance use disorders: Reinforcement pathologies and their repair. *Annual Review of Clinical Psychology*, *10*, 641–677. https://doi.org/10.1146/annurev-clinpsy-032813-153724

Boyce, T. E., & Geller, E. S. (2001). Applied behavior analysis and occupational safety: The challenge of response maintenance. *Journal of Organizational Behavior Management*, *21*(1), 31–60. https://doi.org/10.1300/J075v21n01_03

Bucklin, B. R., & Dickinson, A. M. (2001). Individual monetary incentives: A review of different types of arrangements between performance and pay. *Journal of Organizational Behavior Management*, *21*(3), 45–137. https://doi.org/10.1300/J075v21n03_03

Bureau of Labor Statistics (2019a). *American time use survey summary* (USDL-19-1003). Bureau of Labor Statistics, U.S. Department of Labor. https://www.bls.gov/news.release/atus.nr0.htm

Bureau of Labor Statistics (2019b). *Work experience of the population* (USDL-19-2103). Bureau of Labor Statistics, U.S. Department of Labor. https://www.bls.gov/news.release/work.htm

Call, N. A., Trosclair-Lasserre, N. M., Findley, A. J., Reavis, A. R., & Shillingsburg, M. A. (2012). Correspondence between single versus daily preference assessment outcomes and reinforcer efficacy under progressive-ratio schedules. *Journal of Applied Behavior Analysis*, *45*(4), 763–777. https://doi.org/10.1901/jaba.2012.45-763

Catania, C. A. (2013). *Learning* (5th ed.). Sloan Publishing.

Critchfield, T. S., Tucker, J. A., & Vuchinich, R. E. (1998). Self-report methods. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 435–470). Plenum Press. https://doi.org/10.1007/978-1-4899-1947-2_14

Daniels, A. C., & Bailey, J. S. (2014). *Performance management: Changing behavior that drives organizational performance* (5th ed.). Performance Management Publications.

DeLeon, I. G., Bullock, C. E., & Catania, A. C. (2013). Arranging reinforcement contingencies in applied settings: Fundamentals and implications of recent basic and applied research. In G. J. Madden, W. V. Dube, T. D. Hackenberg, G. P. Hanley, & K. A. Lattal (Eds.), *APA handbook of behavior analysis, Vol. 2. Translating principles into practice* (pp. 47–75). American Psychological Association. https://doi.org/10.1037/13938-003

DeLeon, I. G., Frank, M. A., Gregory, M. K., & Allman, M. J. (2009). On the correspondence between preference assessment outcomes and progressive-ratio schedule assessments of stimulus value. *Journal of Applied Behavior Analysis*, *42*(3), 729–733. https://doi.org/10.1901/jaba.2009.42-729

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. https://doi.org/10.1901/jaba.1996.29-519

DeLeon, I. G., Iwata, B. A., Goh, H.-L., & Worsdell, A. S. (1997). Emergence of reinforcer preference as a function of schedule requirements and similarity. *Journal of Applied Behavior Analysis*, *30*(3), 439–449. https://doi.org/10.1901/jaba.1997.30-439

DeRosa, N. M., & Roane, H. S. (2015). Preference and demand characteristics of reinforcement: practical extensions. In F. D. DiGennaro Reed & D. D. Reed (Eds.), *Autism service delivery*: Bridging the gap between science and practice (pp. 307–325). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-2656-5_11

Diener, L. H., McGee, H. M., & Miguel, C. F. (2009). An integrated approach for conducting a behavioral systems analysis. *Journal of Organizational Behavior Management*, *29*(2), 108–135. https://doi.org/10.1080/01608060902874534

DiGennaro Reed, F. D., Hirst, J. M., & Howard, V. J. (2013). Empirically supported staff selection, training, and management strategies. In D. D. Reed, F. D. DiGennaro Reed, & J. K. Luiselli (Eds.), *Handbook of crisis intervention and developmental disabilities* (pp. 71–85). Springer Science + Business Media. https://doi.org/10.1007/978-1-4614-6531-7_5

Dixon, M. R., & Belisle, J. (2018). Pay for performance: The complex industrial token economy. In B. Wine & J. Pritchard (Eds.), *Organizational behavior management: The essentials*. Hedgehog Publishers.

Fair Labor Standards Act of 1938, 29 U.S.C. §201 *et seq.*

Feiner, I., & Hornik, K. (2019). tm: Text mining package (R package version 0.7-7) [Computer software manual]. https://cran.r-project.org/package=tm

Fisher, W. W., Piazza, C. C., Bowman, L. G., & Amari, A. (1996). Integrating caregiver report with a systematic choice assessment to enhance reinforcer identification. *American Journal on Mental Retardation*, *101*(1), 15–25.

Fisher, W. 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*(2), 491–498. https://doi.org/10.1901/jaba.1992.25-491

Frank-Crawford, M. A., Castillo, M. I., & DeLeon, I. G. (2018). Does preference rank predict substitution for the reinforcer for problem behavior? A behavioral economic analysis. *Journal of Applied Behavior Analysis*, *51*(2), 276–282. https://doi.org/10.1901/jaba.2009.42-729

Frederiksen, L. W., & Lovett, S. B. (1980). Inside organizational behavior management: Perspectives on an emerging field. *Journal of Organizational Behavior Management*, *2*(3), 193–203. https://doi.org/10.1300/J075v02n03_05

Friman, P. C. (2010). Come on in, the water is fine: Achieving mainstream relevance through integration with primary medical care. *The Behavior Analyst*, *33*(1), 19–36. https://doi.org/10.1007/BF03392201

Gentile, N. D., Librizzi, E. H., & Martinetti, M. P. (2012). Academic constraints on alcohol consumption in college students: A behavioral economic analysis. *Experimental and Clinical Psychopharmacology*, *20*(5), 390–399. https://doi.org/10.1037/a0029665

Gilroy, S. P., Kaplan, B. A., & Leader, G. (2018). A systematic review of applied behavioral economics in assessments and treatments for individuals with developmental disabilities. *Review Journal of Autism and Developmental Disorders*, *5*(3), 247–259. https://doi.org/10.1007/s40489-018-0136-6

Hagopian, L. P., Long, E. S., & Rush, K. S. (2004). Preference assessment procedures for individuals with developmental disabilities. *Behavior Modification*, *28*(5), 668–677. https://doi.org/10.1177/0145445503259836

Hanley, G. P., Iwata, B. A., & Roscoe, E. M. (2006). Some determinants of changes in preference over time. *Journal of Applied Behavior Analysis*, *39*(2), 189–202. https://doi.org/10.1901/jaba.2006.163-04

Henley, A. J. (2017). *Translating reinforcer dimensions and behavioral economic demand to inform incentive delivery in organizational behavior management* [Unpublished doctoral dissertation]. University of Kansas. http://hdl.handle.net/1808/26327

Henley, A. J., DiGennaro Reed, F. D., Reed, D. D., & Kaplan, B. A. (2016). A crowdsourced nickel-and-dime approach to analog OBM research: A behavioral economic framework for understanding workforce attrition. *Journal of the Experimental Analysis of Behavior*, *106*(2), 134–144. https://doi.org/10.1002/jeab.220

Henley, A. J., Hirst, J. M., DiGennaro Reed, F. D., Becirevic, A., & Reed, D. D. (2017). Function-altering effects of rule phrasing in the modulation of instructional control. *The Analysis of Verbal Behavior*, *33*(1), 24–40. https://doi.org/10.1007/s40616-016-0063-5

Hodos, W. (1961). Progressive ratio as a measure of reward strength. *Science*, *134*(3483), 943–944. https://doi.org/10.1126/science.134.3483.943

Hursh, S. R. (1980). Economic concepts for the analysis of behavior. *Journal of the Experimental Analysis of Behavior*, *34*(2), 219–238. https://doi.org/10.1901/jeab.1980.34-219

Hursh, S. R. (1984). Behavioral economics. *Journal of the Experimental Analysis of Behavior*, *42*(3), 435–452. https://doi.org/10.1901/jeab.1984.42-435

Hursh, S. R., Madden, G. J., Spiga, R., DeLeon, I. G., & Francisco, M. T. (2013). The translational utility of behavioral economics: The experimental analysis of consumption and choice. In G. J. Madden, W. V. Dube, T. D. Hackenberg, G. P. Hanley, & K. A. Lattal (Eds.), *APA handbook of behavior analysis: Vol. 2. Translating principles into practice* (pp. 191–224). American Psychological Association. https://doi.org/10.1037/13938-008

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. https://doi.org/10.1037/1064-1297.7.4.412

Jarmolowicz, D. P., & Lattal, K. A. (2010). On distinguishing progressively increasing response requirements for reinforcement. *The Behavior Analyst*, *33*(1), 119–125. https://doi.org/10.1007/BF03392207

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). Routledge.

Kang, S., O'Reilly, M., Lancioni, G., Falcomata, T. S., Sigafoos, J., & Xu, Z. (2013). Comparison of the predictive validity and consistency among preference assessment procedures: A review of the literature. *Research in Developmental Disabilities*, *34*(4), 1125–1133. https://doi.org/10.1016/j.ridd.2012.12.021

Kaplan, B. A., Foster, R. N. S., Reed, D. D., Amlung, M., Murphy, J. G., & MacKillop, J. (2018). Understanding alcohol motivation using the alcohol purchase task: A methodological systematic review. *Drug and Alcohol Dependence*, *191*, 117–140. https://doi.org/10.1016/j.drugalcdep.2018.06.029

Kodak, T., Lerman, D. C., & Call, N. (2007). Evaluating the influence of postsession reinforcement on choice of reinforcers. *Journal of Applied Behavior Analysis*, *40*(3), 515–527. https://doi.org/10.1901/jaba.2007.40-515

Kortick, S. A., & O'Brien, R. M. (1996). The world series of quality control: A case study in the package delivery industry. *Journal of Organizational Behavior Management*, *16*(2), 77–93. https://doi.org/10.1300/J075v16n02_05

Lang, D. & Chien, G.-T. (2018). wordcloud2: Create word cloud by 'htmlwidget' (R package version 0.2.1) [Computer software manual]. https://cran.r-project.org/package=wordcloud2

Lea, S. E. G. (1978). The psychology and economics of demand. *Psychological Bulletin*, *85*(3), 441–466. https://doi.org/10.1037/0033-2909.85.3.441

Lee, M. S., Yu, C. T., Martin, T. L., & Martin, G. L. (2010). On the relation between reinforcer efficacy and preference. *Journal of Applied Behavior Analysis*, *43*(1), 95–100. https://doi.org/10.1901/jaba.2010.43-95

Lüdecke, D. (2018). *sjmisc: Data and variable transformation functions* (R package version 2.8.2) [Computer software manual]. https://CRAN.R-project.org/package=sjmisc

Madden, G. J., & Perone, M. (1999). Human sensitivity to concurrent schedules of reinforcement: Effects of observing schedule-correlated stimuli. *Journal of the Experimental Analysis of Behavior*, *71*(3), 303–318. https://doi.org/10.1901/jeab.1999.71-303

Mazur, J. E. (1993). Predicting the strength of a conditioned reinforcer: Effects of delay and uncertainty. *Current Directions in Psychological Science*, *2*(3), 70–74. https://doi.org/10.1111/1467-8721.ep10770907

McLeod, A. I. (2011). *Kendall: Kendall rank correlation and Mann–Kendall trend test* (R package version 2.2) [Computer software manual]. https://CRAN.R-project.org/package=Kendall

Moher D., Liberati A., Tetzlaff J., & Altman D. G., The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *PLoS Med*, *6*(7): e1000097. https://doi.org/10.1371/journal.pmed.1000097

Murphy, J. G., & MacKillop, J. (2006). Relative reinforcing efficacy of alcohol among college student drinkers. *Experimental and Clinical Psychopharmacology*, *14*(2), 219–227. https://doi.org/10.1037/1064-1297.14.2.219

Neef, N. A., Shade, D., & Miller, M. S. (1994). Assessing influential dimensions of reinforcers on choice in students with serious emotional disturbance. *Journal of Applied Behavior Analysis*, *27*(4), 575–583. https://doi.org/10.1901/jaba.1994.27-575

Northup, J. (2000). Further evaluation of the accuracy of reinforcer surveys: A systematic replication. *Journal of Applied Behavior Analysis*, *33*(3), 335–338. https://doi.org/10.1901/jaba.2000.33-335

Northup, J., George, T., Jones, K., Broussard, C., & Vollmer, T. R. (1996). A comparison of reinforcer assessment methods: The utility of verbal and pictorial choice procedures. *Journal of Applied Behavior Analysis*, *29*(2), 201–212. https://doi.org/10.1901/jaba.1996.29-201

Novak, M. D., DiGennaro Reed, F. D., Erath, T. G., Blackman, A. L., Ruby, S. A. & Pellegrino, A. J. (2019). Evidence-based performance management: Applying behavioral science to support practitioners. *Perspectives on Behavior Science*, *42*(4), 955–972. https://doi.org/10.1007/s40614-019-00232-z

Oah, S., & Lee, J.-H. (2011). Effects of hourly, low-incentive, and high-incentive pay on simulated work productivity: Initial findings with a new laboratory method. *Journal of Organizational Behavior Management*, *31*(1), 21–42. https://doi.org/10.1080/01608061.2011.541820

Pace, G. M., Ivancic, M. T., Edwards, G. L., Iwata, B. A., & Page, T. J. (1985). Assessment of stimulus preference and reinforcer value with profoundly retarded individuals. *Journal of Applied Behavior Analysis*, *18*(3), 249–255. https://doi.org/10.1901/jaba.1985.18-249

Parsons, M. B., Rollyson, J. H., & Reid, D. H. (2012). Evidence-based staff training: A guide for practitioners. *Behavior Analysis in Practice*, *5*(2), 2–11. https://doi.org/10.1007/BF03391819

Penrod, B., Wallace, M. D., & Dyer, E. J. (2008). Assessing potency of high- and low-preference reinforcers with respect to response rate and response patterns. *Journal of Applied Behavior Analysis*, *41*(2), 177–188. https://doi.org/10.1901/jaba.2008.41-177

Peterson, C., Lerman, D. C., & Nissen, M. A. (2016). Reinforcer choice as an antecedent versus consequence. *Journal of Applied Behavior Analysis*, *49*(2), 286–293. https://doi.org/10.1002/jaba.284

Petry, N. M., & Bickel, W. K. (1998). Polydrug abuse in heroin addicts: A behavioral economic analysis. *Addiction*, *93*(3), 321–335. https://doi.org/10.1046/j.1360-0443.1998.9333212.x

Piazza, C. C., Fisher, W. W., Hagopian, L. P., Bowman, L. G., & Toole, L. (1996). Using a choice assessment to predict reinforcer effectiveness. *Journal of Applied Behavior Analysis*, *29*(1), 1–9. https://doi.org/10.1901/jaba.1996.29-1

Piazza, C. C., Roane, H. S., & Karsten, A. (2011). Identifying and enhancing the effectiveness of positive reinforcement. In W. W. Fisher, C. C. Piazza, & H. S. Roane (Eds.), *Handbook of applied behavior analysis* (pp. 151–164). The Guilford Press.

Poling, A., Dickinson, A. M., Austin, J., & Normand, M. (2000). Basic behavioral research and organizational behavior management. In J. Austin & J. E. Carr (Eds.), *Handbook of applied behavior analysis* (pp. 295–320). Context Press.

R Core Team (2019). *R: A language and environment for statistical computing* (Version 3.6.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Reed, D. D., DiGennaro Reed, F. D., Campisano, N., Lacourse, K., & Azulay, R. L. (2012). Assessing and increasing staff preference for job tasks using concurrent-chains schedules and probabilistic outcomes. *Journal of Organizational Behavior Management*, *32*(3), 253–262. https://doi.org/10.1080/01608061.2012.698121

Reed, D. D., Kaplan, B. A., & Becirevic, A. (2015). Basic research on the behavioral economics of reinforcer value. In F. D. DiGennaro Reed & D. D. Reed (Eds.), *Autism service delivery: Bridging the gap between science and practice* (pp. 279–306). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-2656-5_10

Reed, D. D., Luiselli, J. K., Magnuson, J. D., Fillers, S., Vieira, S., & Rue, H. C. (2009). A comparison between traditional economical and demand curve analyses of relative reinforcer efficacy in the validation of preference assessment predictions. *Developmental Neurorehabilitation*, *12*(3), 164–169. https://doi.org/10.1080/17518420902858983

Reid, D. H., & Parsons, M. B. (2000). Organizational behavior management in human service settings. In J. Austin & J. E. Carr (Eds.), *Handbook of applied behavior analysis* (pp. 275–294). Context Press.

Roane, H. S. (2008). On the applied use of progressive-ratio schedules of reinforcement. *Journal of Applied Behavior Analysis*, *41*(2), 155–161. https://doi.org/10.1901/jaba.2008.41-155

Roane, H. S., Call, N. A., & Falcomata, T. S. (2005). A preliminary analysis of adaptive responding under open and closed economies. *Journal of Applied Behavior Analysis*, *38*(3), 335–348. https://doi.org/10.1901/jaba.2005.85-04

Roane, H. S., Lerman, D. C., & Vorndran, C. M. (2001). Assessing reinforcers under progressive schedule requirements. *Journal of Applied Behavior Analysis*, *34*(2), 145–167. https://doi.org/10.1901/jaba.2001.34-145

Roma, P. G., Hursh, S. R., & Hudja, S. (2015). Hypothetical purchase task questionnaires for behavioral economic assessments of value and motivation. *Managerial and Decision Economics*, *37*(4–5), 306–323. https://doi.org/10.1002/mde.2718

Roma, P. G., Reed, D. D., DiGennaro Reed, F. D., & Hursh, S. R. (2017). Progress of and prospects for hypothetical purchase task questionnaires in consumer behavior analysis and public policy. *The Behavior Analyst*, *40*(2), 329–342. https://doi.org/10.1007/s40614-017-0100-2

Roscoe, E. M., Iwata, B. A., & Kahng, S. (1999). Relative versus absolute reinforcement effects: Implications for preference assessments. *Journal of Applied Behavior Analysis*, *32*(4), 479–493. https://doi.org/10.1901/jaba.1999.32-479

Siegal, S. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw–Hill Book Company.

Sigurdsson, S. O., & Austin, J. (2006). Institutionalization and response maintenance in organizational behavior management. *Journal of Organizational Behavior Management*, *26*(4), 41–77. https://doi.org/10.1300/J075v26n04_03

Skidmore, J. R., & Murphy, J. G. (2011). The effect of drink price and next-day responsibilities on college student drinking: A behavioral economic analysis. *Psychology of Addictive Behaviors*, *25*(1), 57–68. https://doi.org/10.1037/a0021118

Taravella, C. C., Lerman, D. C., Contrucci, S. A., & Roane, H. S. (2000). Further evaluation of low-ranked items in stimulus-choice preference assessments. *Journal of Applied Behavior Analysis*, *33*(1), 105–108. https://doi.org/10.1901/jaba.2000.33-105

Tiger, J. H., Toussaint, K. A., & Roath, C. T. (2010). An evaluation of the value of choice-making opportunities in single-operant arrangements: Simple fixed- and progressive-ratio schedules. *Journal of Applied Behavior Analysis*, *43*(3), 519–524. https://doi.org/10.1901/jaba.2010.43-519

Timberlake, W., & Peden, B. F. (1987). On the distinction between open and closed economies. *Journal of the Experimental Analysis of Behavior*, *48*(1), 35–60. https://doi.org/10.1901/jeab.1987.48-35

Tustin, R. D. (1994). Preference for reinforcers under varying schedule arrangements: A behavioral economic analysis. *Journal of Applied Behavior Analysis*, *27*(4), 597–606. https://doi.org/10.1901/jaba.1994.27-597

VandenBos, G. R. (Ed.). (2015). *APA dictionary of psychology* (2nd ed.). American Psychological Association. https://doi.org/10.1037/14646-000

VanStelle, S. E., Vicars, S. M., Harr, V., Miguel, C. F., Koerber, J. L., Kazbour, R., & Austin, J. (2012). The publication history of *Journal of Organizational Behavior Management*: An objective review and analysis: 1998–2009. *Journal of Organizational Behavior Management*, *32*(2), 93–123. https://doi.org/10.1080/01608061.2012.675864

Virués-Ortega, J., Pritchard, K., Grant, R. L., North, S., Hurtado-Parrado, C., Lee, M. S. H., Temple, B., Julio, F., & Yu, C. T. (2014). Clinical decision making and preference assessment for individuals with intellectual and developmental disabilities. *American*

Journal on Intellectual and Developmental Disabilities, *119*(2), 151–170. https://doi.org/10.1352/1944-7558-119.2.151

Wacker, D. P., Berg, W. K., Wiggins, B., Muldoon, M., & Cavanaugh, J. (1985). Evaluation of reinforcer preferences for profoundly handicapped students. *Journal of Applied Behavior Analysis*, *18*(2), 173–178. https://doi.org/10.1901/jaba.1985.18-173

Waldvogel, J. M., & Dixon, M. R. (2008). Exploring the utility of preference assessments in organizational behavior management. *Journal of Organizational Behavior Management*, *28*(1), 76–87. https://doi.org/10.1080/01608060802006831

Wilder, D. A., Austin, J., & Casella, S. (2009). Applying behavior analysis in organizations: Organizational behavior management. *Psychological Services*, *6*(3), 202–211. https://doi.org/10.1037/a0015393

Wilder, D. A., Harris, C., Casella, S., Wine, B., & Postma, N. (2011). Further evaluation of the accuracy of managerial prediction of employee preference. *Journal of Organizational Behavior Management*, *31*(2), 130–139. https://doi.org/10.1080/01608061.2011.569202

Wilder, D. A., Rost, K., & McMahon, M. (2007). The accuracy of managerial prediction of employee preference: A brief report. *Journal of Organizational Behavior Management*, *27*(2), 1–14. https://doi.org/10.1300/J075v27n02_01

Wilder, D. A., Therrien, K., & Wine, B. (2006). A comparison between survey and verbal choice methods of identifying potential reinforcers among employees. *Journal of Organizational Behavior Management*, *25*(4), 1–13. https://doi.org/10.1300/J075v25n04_01

Wine, B., & Axelrod, S. (2014). The effects of progressively thinning high-preference item delivery on responding in employees. *Journal of Organizational Behavior Management*, *34*(4), 291–299. https://doi.org/10.1080/01608061.2014.973629

Wine, B., Gilroy, S., & Hantula, D. A. (2012). Temporal (in)stability of employee preferences for rewards. *Journal of Organizational Behavior Management*, *32*(1), 58–64. https://doi.org/10.1080/01608061.2012.646854

Wine, B., Gugliemella, C., & Axelrod, S. (2013). An examination of generalized conditioned reinforcers in stimulus preference assessments. *Journal of Organizational Behavior Management*, *33*(4), 244–251. https://doi.org/10.1080/01608061.2013.843433

Wine, B., Kelley, D. P., & Wilder, D. A. (2014). An initial assessment of effective preference assessment intervals among employees. *Journal of Organizational Behavior Management*, *34*(3), 188–195. https://doi.org/10.1080/01608061.2014.944747

Wine, B., Reis, M., & Hantula, D. A. (2014). An evaluation of stimulus preference assessment methodology in organizational behavior management. *Journal of Organizational Behavior Management*, *34*(1), 7–15. https://doi.org/10.1080/01608061.2013.873379

Wine, B., & Wilder, D. A. (2009). The effects of varied versus constant high-, medium, and low-preference stimuli on performance. *Journal of Applied Behavior Analysis*, *42*(2), 321–326. https://doi.org/10.1901/jaba.2009.42-321

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*(2), 203–214. https://doi.org/10.1901/jaba.1978.11-203

Zhou, L., Iwata, B. A., Goff, G. A., & Shore, B. A. (2001). Longitudinal analysis of leisure-item preferences. *Journal of Applied Behavior Analysis*, *34*(2), 179–184. https://doi.org/10.1901/jaba.2001.34-179

Table 1

Participant Demographics and Earnings

Participant	Gender	Age	Hours	Chocolate	Fruity Candy	Gel Pen	Gift Card	Gum	Lip Balm
Julianne	Female	19	3.5	1	1	6	14	0	8
Shannon	Female	19	4.75	7	8	11	12	10	11
Alexia	Female	19	3.75	8	11	0	14	5	8
Cindy	Female	21	3.75	8	9	5	15	9	8
Elise	Female	19	3.75	6	6	0	8	6	4
Marie	Female	23	4.75	12	12	1	12	10	8
	Mean	20	3.5	7.0	8.0	3.8	12.5	6.7	7.8

Table 2

Progressive Ratio Schedule Progressions

	Julianne	Shannon	Alexia	Cindy	Elise	Marie
PR 1	5	15	5	5	5	15
PR 2	11	20	11	15	11	25
PR 3	15	30	15	20	15	35
PR 4	20	35	20	30	20	40
PR 5	30	40	30	35	30	45
PR 6	35	45	35	40	35	50

Table 3

Average Kendall Tau Correlation Coefficients

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	.834	—						
3. HWT	.907	.842	—					
Reinforcer Assessments								
4. FR	.651	.511	.622	—				
5. PR	.293	.285	.398	.652	—			
Follow-up PAs								
6. Survey	.326	.264	.343	.624	.839	—		
7. Ranking	.331	.289	.312	.632	.859	.882	—	
8. HWT	.307	.241	.310	.593	.845	.886	.904	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

Table 4

Kendall Tau Correlation Coefficients: Julianne

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	.645	—						
3. HWT	1.0*	.645	—					
Reinforcer Assessments								
4. FR	.923*	.501	.923*	—				
5. PR	.721	.298	.721	.881*	—			
Follow-up PAs								
6. Survey	.815*	.414	.815*	.964*	.926*	—		
7. Ranking	.788	.467	.788	.931*	.894*	.966*	—	
8. HWT	.769	.358	.769	.923*	.881*	.964*	.931*	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

*p < .05.

Table 5

Kendall Tau Correlation Coefficients: Shannon

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	.856*	—						
3. HWT	.886*	.966*	—					
Reinforcer Assessments								
4. FR	–	–	–	—				
5. PR	.081	.276	.286	–	—			
Follow-up PAs								
6. Survey	.084	.358	.296	–	.889*	—		
7. Ranking	.078	.333	.276	–	.966*	.931*	—	
8. HWT	0.0	.276	.214	–	.929*	.964*	.966*	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

*p < .05.

Table 6

Kendall Tau Correlation Coefficients: Alexia

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	.966*	—						
3. HWT	.964*	.931*	—					
Reinforcer Assessments								
4. FR	.624	.602	.555	—				
5. PR	.357	.414	.296	.802	—			
Follow-up PAs								
6. Survey	.322	.389	.251	.804	.806	—		
7. Ranking	.276	.333	.215	.775	.966*	.856*	—	
8. HWT	.089	.086	0.0	.667	.802	.704	.775	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

*$p < .05$.

Table 7

Kendall Tau Correlation Coefficients: Cindy

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	.856*	—						
3. HWT	.806	.690	—					
Reinforcer Assessments								
4. FR	.674	.577	.598	—				
5. PR	.242	.276	.500	.598	—			
Follow-up PAs								
6. Survey	.251	.215	.445	.620	.964*	—		
7. Ranking	.389	.333	.276	.577	.690	.788	—	
8. HWT	.364	.389	.403	.539	.806	.836	.856*	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

*$p < .05$.

Table 8

Kendall Tau Correlation Coefficients: Elise (Reordered)

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	.856*	—						
3. HWT	1.0*	.856*	—					
Reinforcer Assessments								
4. FR	.435	.298	.435	—				
5. PR	−.261	−.298	−.261	.333	—			
Follow-up PAs								
6. Survey	−.364	−.389	−.364	.087	.783	—		
7. Ranking	−.234	−.200	−.234	.298	.894*	.856*	—	
8. HWT	−.167	−.215	−.167	.240	.881*	.920*	.931*	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

*p < .05.

Table 9

Kendall Tau Correlation Coefficients: Marie

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	.828*	—						
3. HWT	.786	.966*	—					
Reinforcer Assessments								
4. FR	.598	.577	.598	—				
5. PR	.617	.745	.849*	.645	—			
Follow-up PAs								
6. Survey	.849*	.596	.617	.645	.667	—		
7. Ranking	.690	.467	.552	.577	.745	.894*	—	
8. HWT	.786	.552	.643	.598	.772	.926*	.966*	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

*$p < .05$.

Table 10

Kendall Tau Correlation Coefficients: Elise (Original)

	1	2	3	4	5	6	7	8
Preference Assessments (PAs)								
1. Survey	—							
2. Ranking	−.856*	—						
3. HWT	1.0*	−.856*	—					
Reinforcer Assessments								
4. FR	.435	−.298	.435	—				
5. PR	−.261	.298	−.261	.333	—			
Follow-up PAs								
6. Survey	−.364	.389	−.364	.087	.783	—		
7. Ranking	−.234	.200	−.234	.298	.894*	.856*	—	
8. HWT	−.167	.215	−.167	.240	.881*	.920*	.931*	—

Note: HWT = hypothetical work task; FR = fixed ratio assessment; PR = progressive ratio assessment. FR values are omitted because no correlation could be calculated.

*p < .05.

Table 11

Preference Assessment Administration Durations (in seconds)

		Survey		Ranking		HWT	
	n	*M*	Range	*M*	Range	*M*	Range
Initial	6	27.0	17.3–40.4	38.3	27.4–50.7	68.4	46.1–97.7
Follow-up	6	27.2	21.3–33.1	21.1	15.9–27.1	58.8	36.6–89.6
All		27.1		29.7		63.6	

Note: HWT = hypothetical work task

Table 12

Preference Assessment Accuracy

Participant	Efficacious Reinforcer True Positive	Efficacious Reinforcer False Negative	Inefficacious Reinforcer True Negative	Inefficacious Reinforcer False Positive	Accuracy
			Survey		
Julianne	1.00 (3)	0.00 (0)	1.00 (3)	0.00 (0)	100.0%
Shannon	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Alexia	0.75 (3)	0.25 (1)	1.00 (2)	0.00 (0)	83.3%
Cindy	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Elise	1.00 (4)	0.00 (0)	1.00 (2)	0.00 (0)	100.0%
Marie	0.80 (4)	0.20 (1)	1.00 (1)	0.00 (0)	83.3%
Mean	0.76 (20)	0.24 (8)	1.00 (8)	0.00 (0)	77.8%
			Ranking (Reordered)		
Julianne	0.67 (2)	0.33 (1)	0.67 (2)	0.33 (1)	66.7%
Shannon	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Alexia	0.75 (3)	0.25 (1)	1.00 (2)	0.00 (0)	83.3%
Cindy	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Elise	0.75 (3)	0.25 (1)	1.00 (2)	0.00 (0)	83.3%
Marie	0.60 (3)	0.40 (2)	1.00 (1)	0.00 (0)	66.7%
Mean	0.63 (17)	0.37 (11)	0.92 (7)	0.08 (1)	66.7%
			HWT		
Julianne	1.00 (3)	0.00 (0)	1.00 (3)	0.00 (0)	100.0%
Shannon	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Alexia	0.75 (3)	0.25 (1)	1.00 (2)	0.00 (0)	83.3%
Cindy	0.67 (4)	0.33 (2)	– (0)	– (0)	66.7%
Elise	1.00 (4)	0.00 (0)	1.00 (2)	0.00 (0)	100.0%
Marie	0.80 (4)	0.20 (1)	1.00 (1)	0.00 (0)	83.3%
Mean	0.79 (21)	0.21 (7)	1.00 (8)	0.00 (0)	80.6%
			Ranking (Original)		
Elise	0.25 (1)	0.75 (3)	0.00 (0)	1.00 (2)	16.7%
Mean	0.54 (15)	0.46 (13)	0.67 (5)	0.33 (3)	55.6%

Note: Proportion of efficacious and inefficacious reinforcers that were classified as HP (true or false positive) or LP (true or false negative). Accuracy is calculated as the sum of true positives and true negatives divided by the number of all outcomes.

Table 13

Follow-up Preference Assessment Accuracy

	Efficacious Reinforcer		Inefficacious Reinforcer		
Participant	True Positive	False Negative	True Negative	False Positive	Accuracy
			Follow-up Survey		
Julianne	1.00 (3)	0.00 (0)	1.00 (3)	0.00 (0)	100.0%
Shannon	0.67 (4)	0.33 (2)	– (0)	– (0)	66.7%
Alexia	0.75 (3)	0.25 (1)	1.00 (2)	0.00 (0)	83.3%
Cindy	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Elise	0.50 (2)	0.50 (2)	0.50 (1)	0.50 (1)	50.0%
Marie	0.60 (3)	0.40 (2)	1.00 (1)	0.00 (0)	66.7%
Mean	0.67 (18)	0.33 (10)	0.88 (7)	0.13 (1)	69.4%
			Follow-up Ranking		
Julianne	1.00 (3)	0.00 (0)	1.00 (3)	0.00 (0)	100.0%
Shannon	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Alexia	0.75 (3)	0.25 (1)	1.00 (2)	0.00 (0)	83.3%
Cindy	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Elise	0.50 (2)	0.50 (2)	0.50 (1)	0.50 (1)	50.0%
Marie	0.60 (3)	0.40 (2)	1.00 (1)	0.00 (0)	66.7%
Mean	0.64 (17)	0.36 (11)	0.88 (7)	0.13 (1)	66.7%
			Follow-up HWT		
Julianne	1.00 (3)	0.00 (0)	1.00 (3)	0.00 (0)	100.0%
Shannon	0.50 (3)	0.50 (3)	– (0)	– (0)	50.0%
Alexia	1.00 (4)	0.00 (0)	0.50 (1)	0.50 (1)	83.3%
Cindy	0.67 (4)	0.33 (2)	– (0)	– (0)	66.7%
Elise	0.50 (2)	0.50 (2)	0.50 (1)	0.50 (1)	50.0%
Marie	0.60 (3)	0.40 (2)	1.00 (1)	0.00 (0)	66.7%
Mean	0.71 (19)	0.29 (9)	0.75 (6)	0.25 (2)	69.4%

Note: Proportion of efficacious and inefficacious reinforcers that were classified as HP (true or false positive) or LP (true or false negative). Accuracy is calculated as the sum of true positives and true negatives divided by the number of all outcomes.

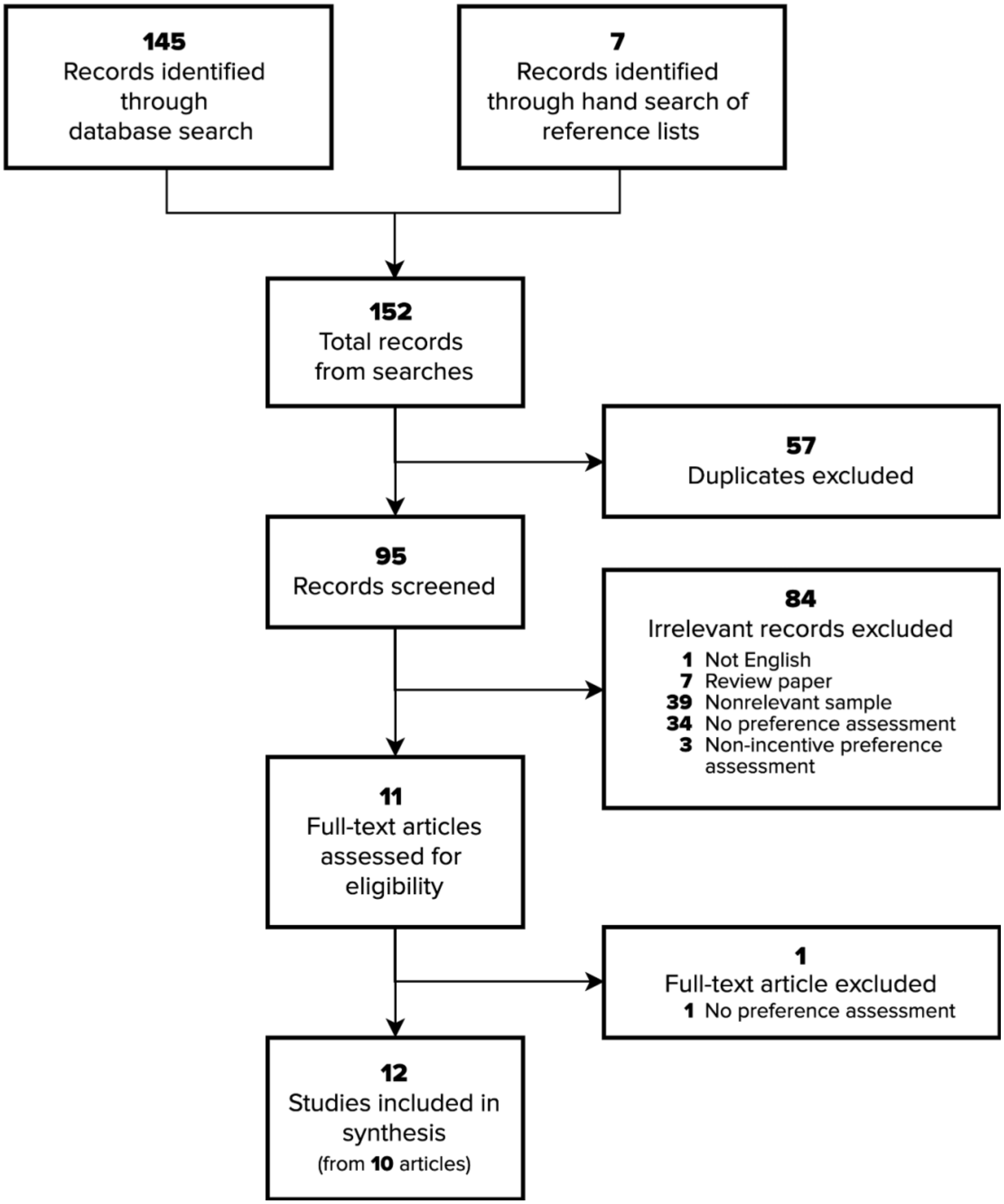
Table 14

Progressive Ratio Assessment Outcomes

Participant	Chocolate	Fruity Candy	Gel Pen	Gift Card	Gum	Lip Balm
			Average Break Point			
Julianne	0.0	0.0	7.0	18.3	0.0	8.3
Shannon	16.7	21.7	26.7	30.0	25.0	26.7
Alexia	8.7	13.7	0.0	18.3	7.3	8.7
Cindy	11.7	13.3	6.7	30.0	15.0	11.7
Elise	3.7	3.7	0.0	12.3	3.7	1.7
Marie	35.0	35.0	5.0	35.0	28.3	21.7
			Cumulative Points			
Julianne	0	0	27	133	0	41
Shannon	65	125	167	206	150	165
Alexia	47	78	0	133	32	47
Cindy	45	65	25	213	60	45
Elise	18	16	0	64	16	5
Marie	226	226	15	225	165	105

Figure 1.

Study-Selection Flow Diagram



Note. Search conducted January 23, 2020.

Figure 2.

Word Cloud of Incentives Used in Reviewed Studies

Note. Font size is proportional to frequency of appearance. Dollar amounts and other units were not included.

Figure 3.

Work Task Interface

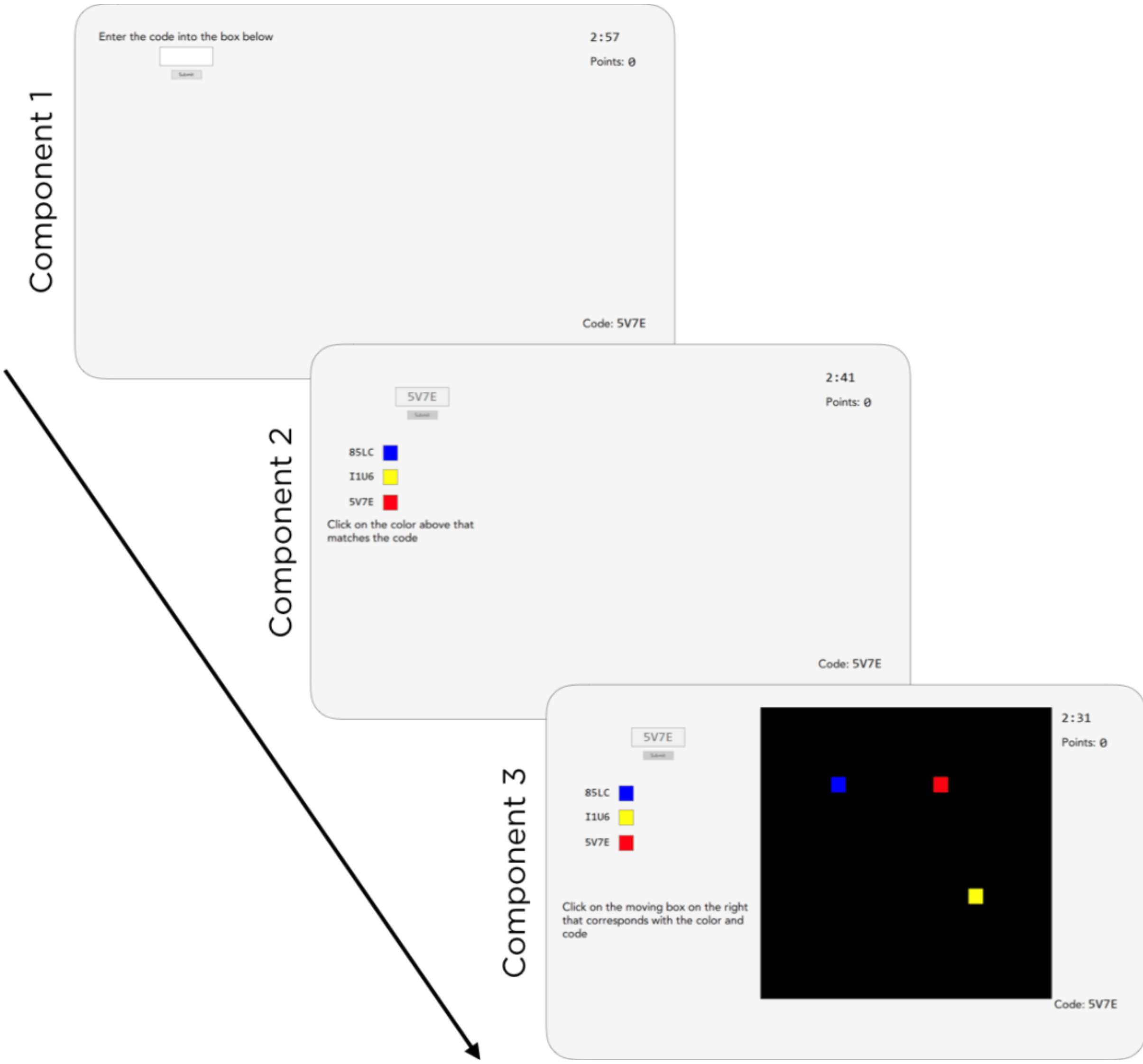
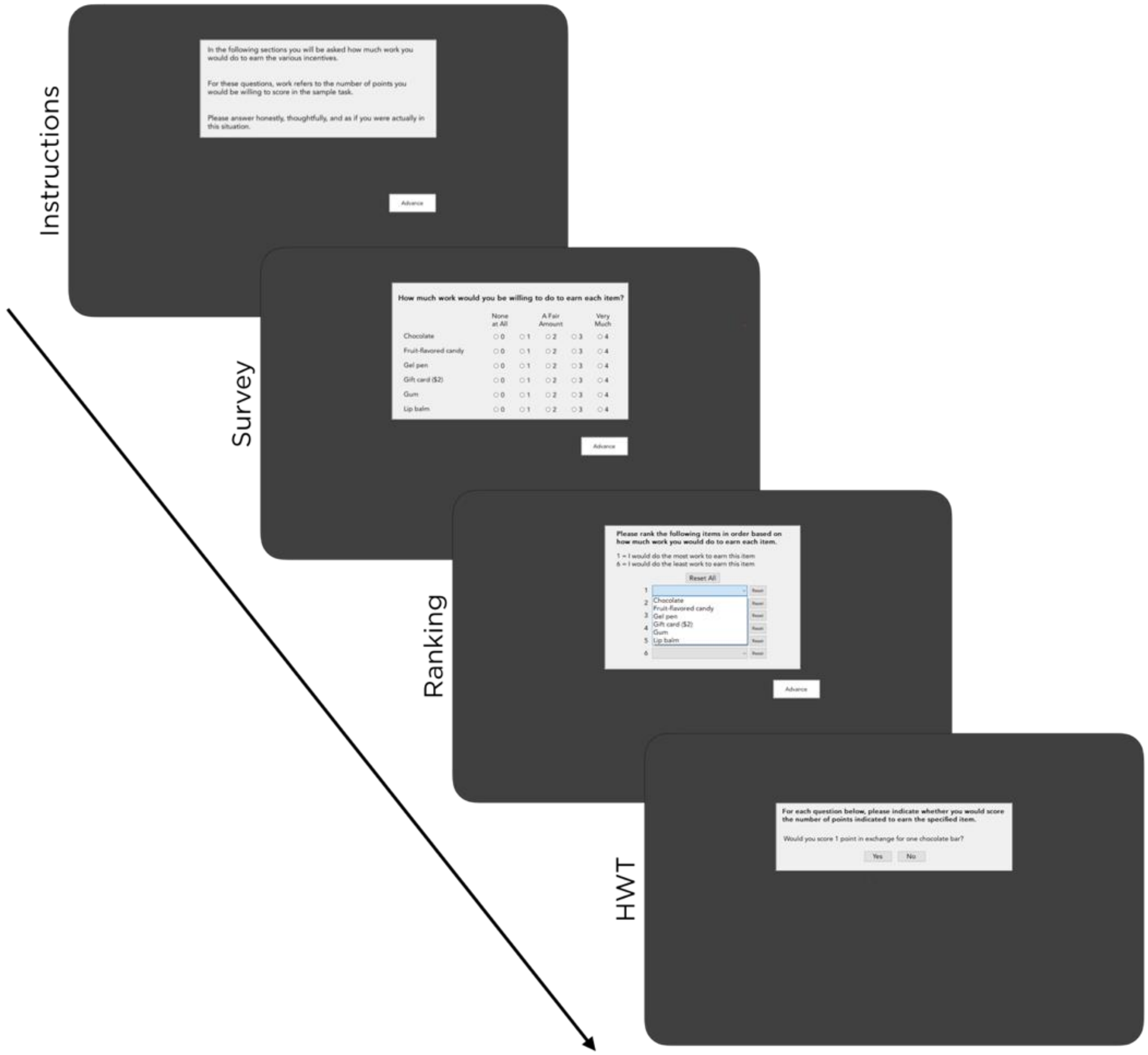


Figure 4.

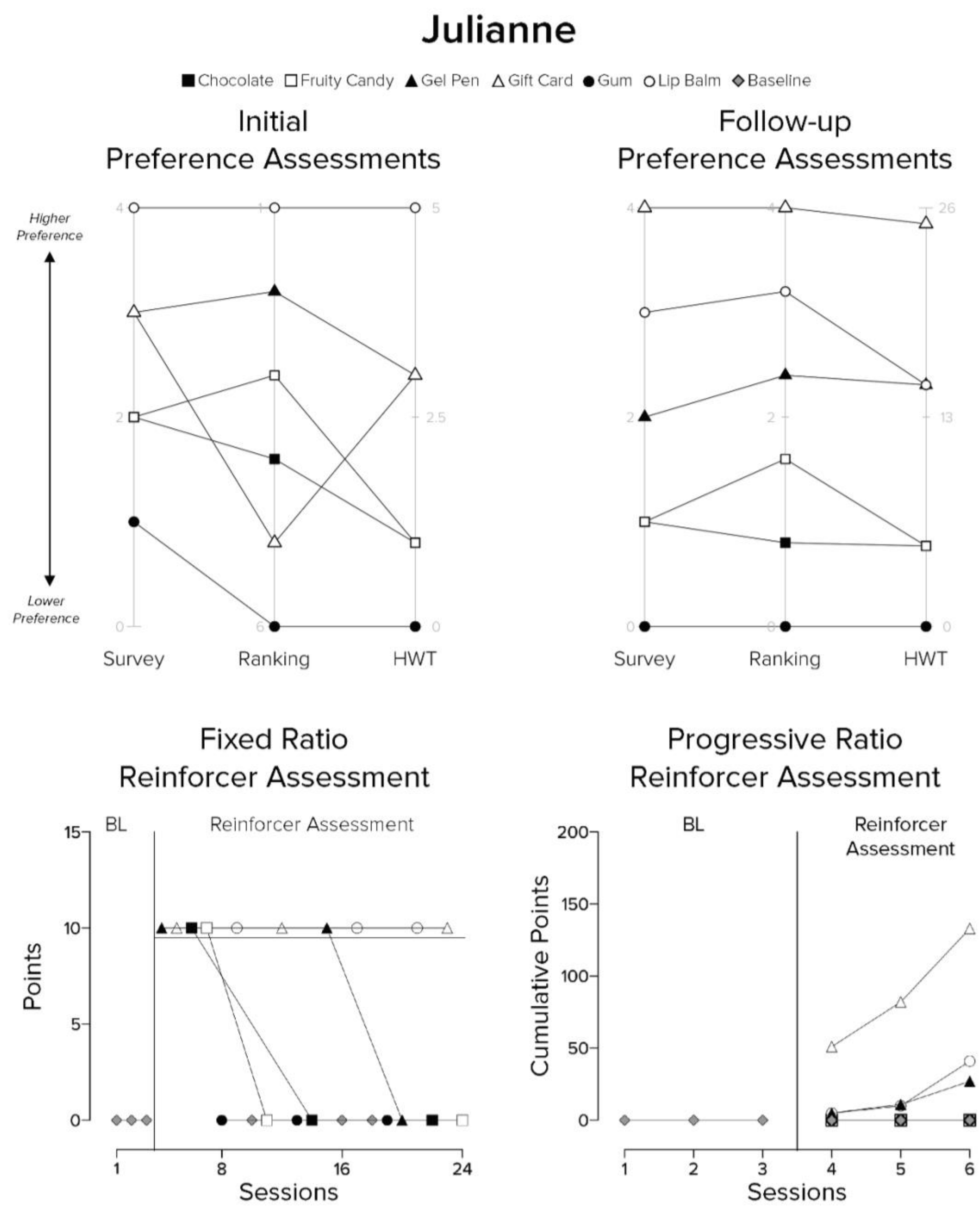
Preference Assessment Interface



Note. HWT = hypothetical work task break point.

Figure 5.

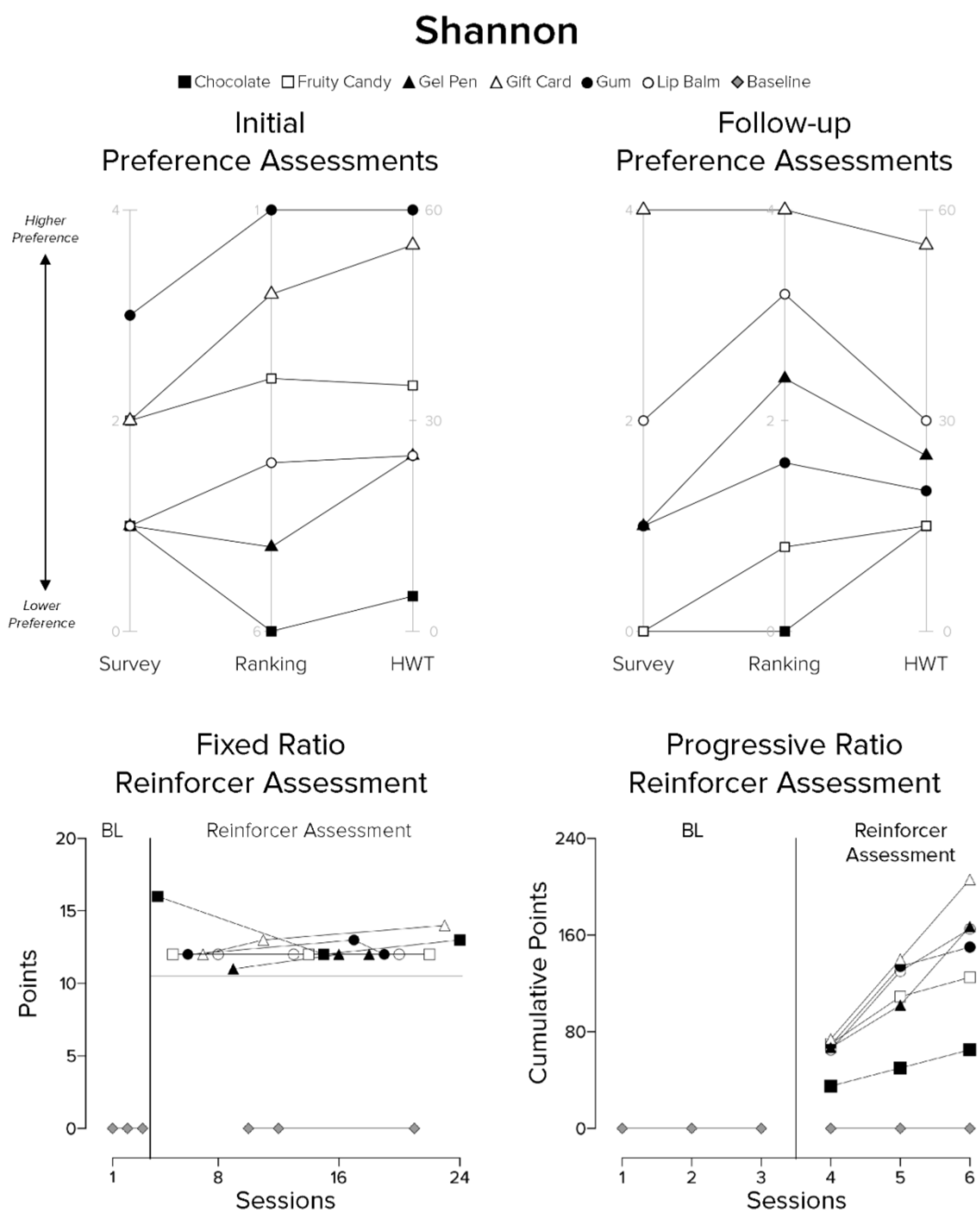
Assessment Data: Julianne



Note. HWT = hypothetical work task break point; BL = baseline

Figure 6.

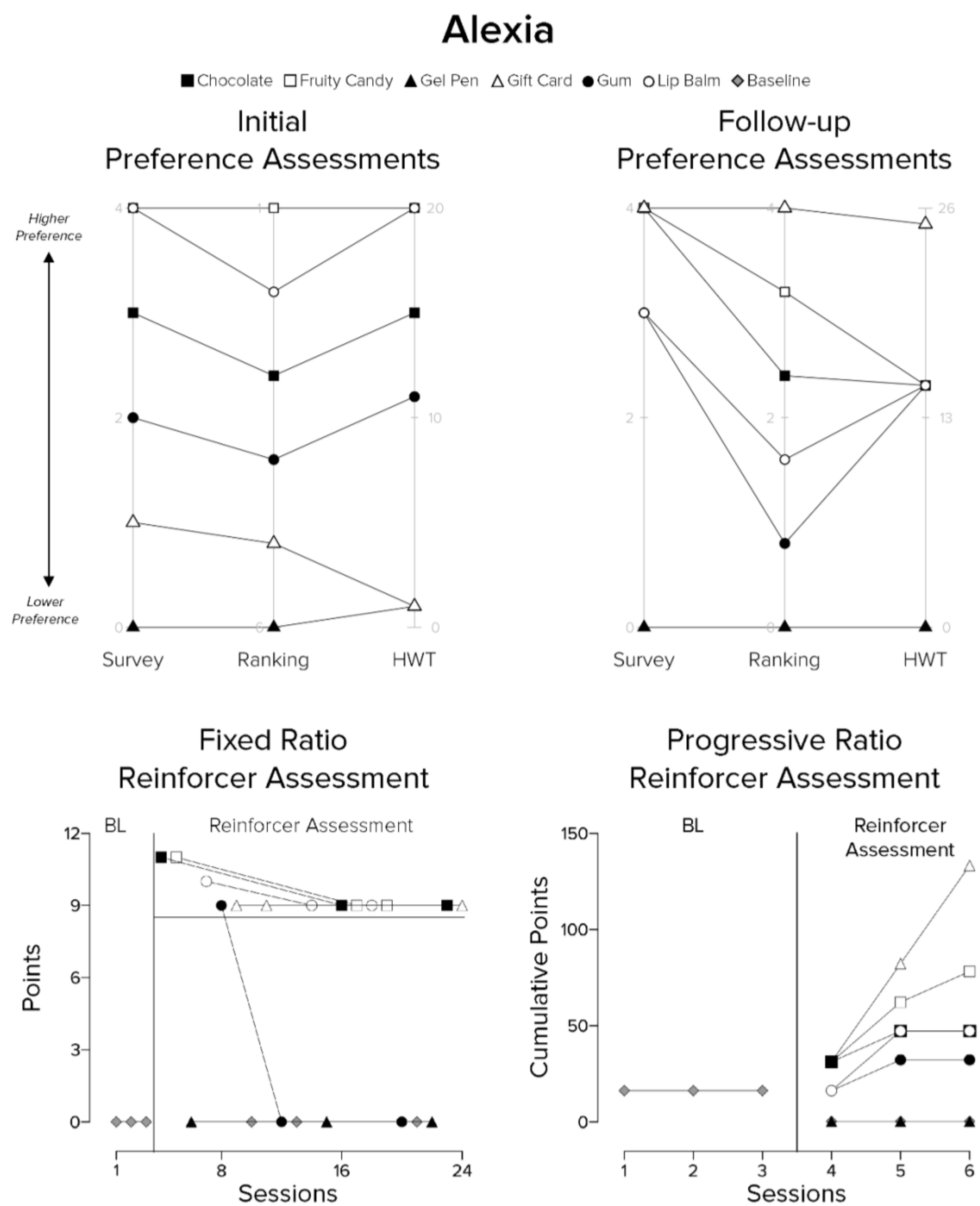
Assessment Data: Shannon



Note. HWT = hypothetical work task break point; BL = baseline

Figure 7.

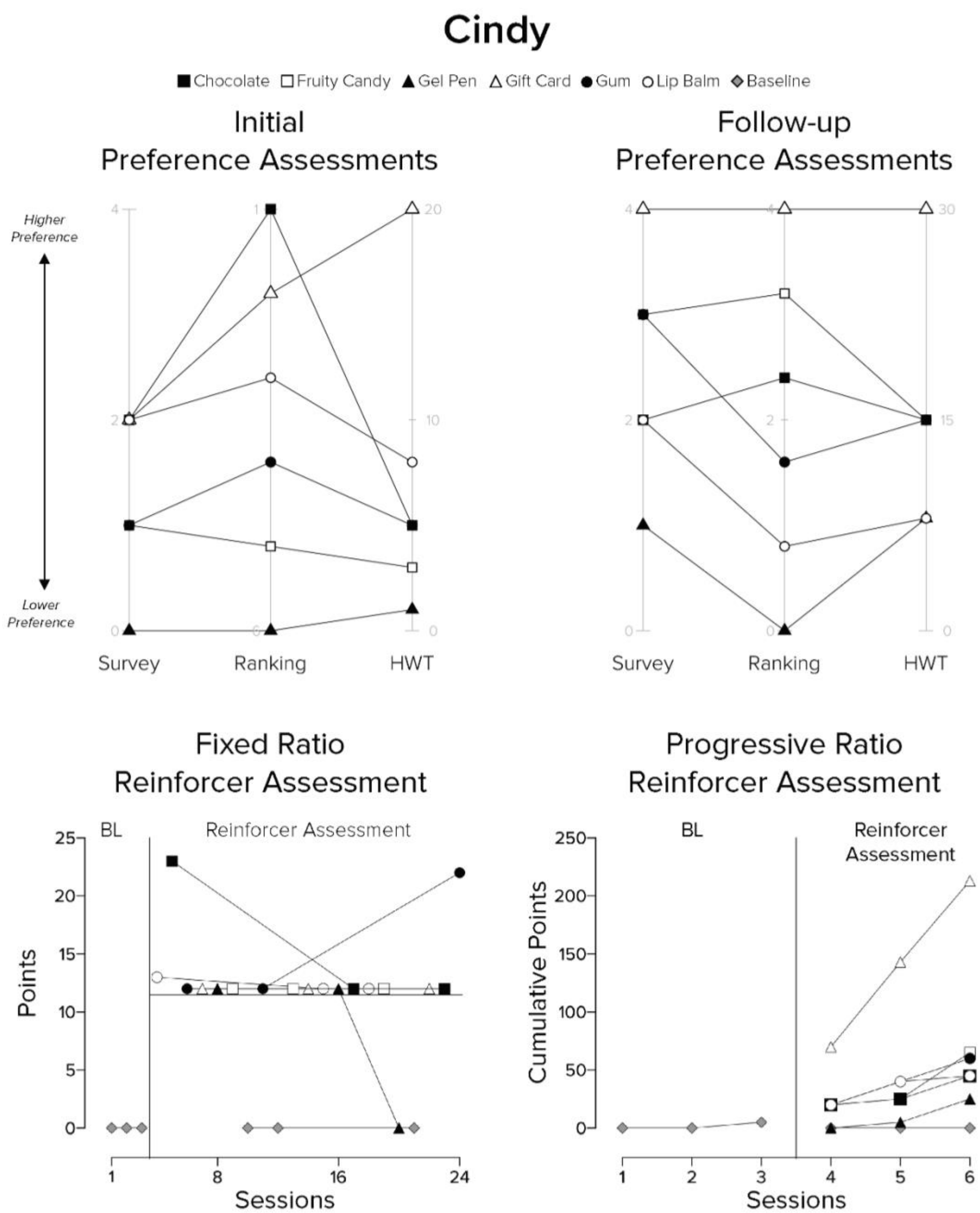
Assessment Data: Alexia



Note. HWT = hypothetical work task break point; BL = baseline

Figure 8.

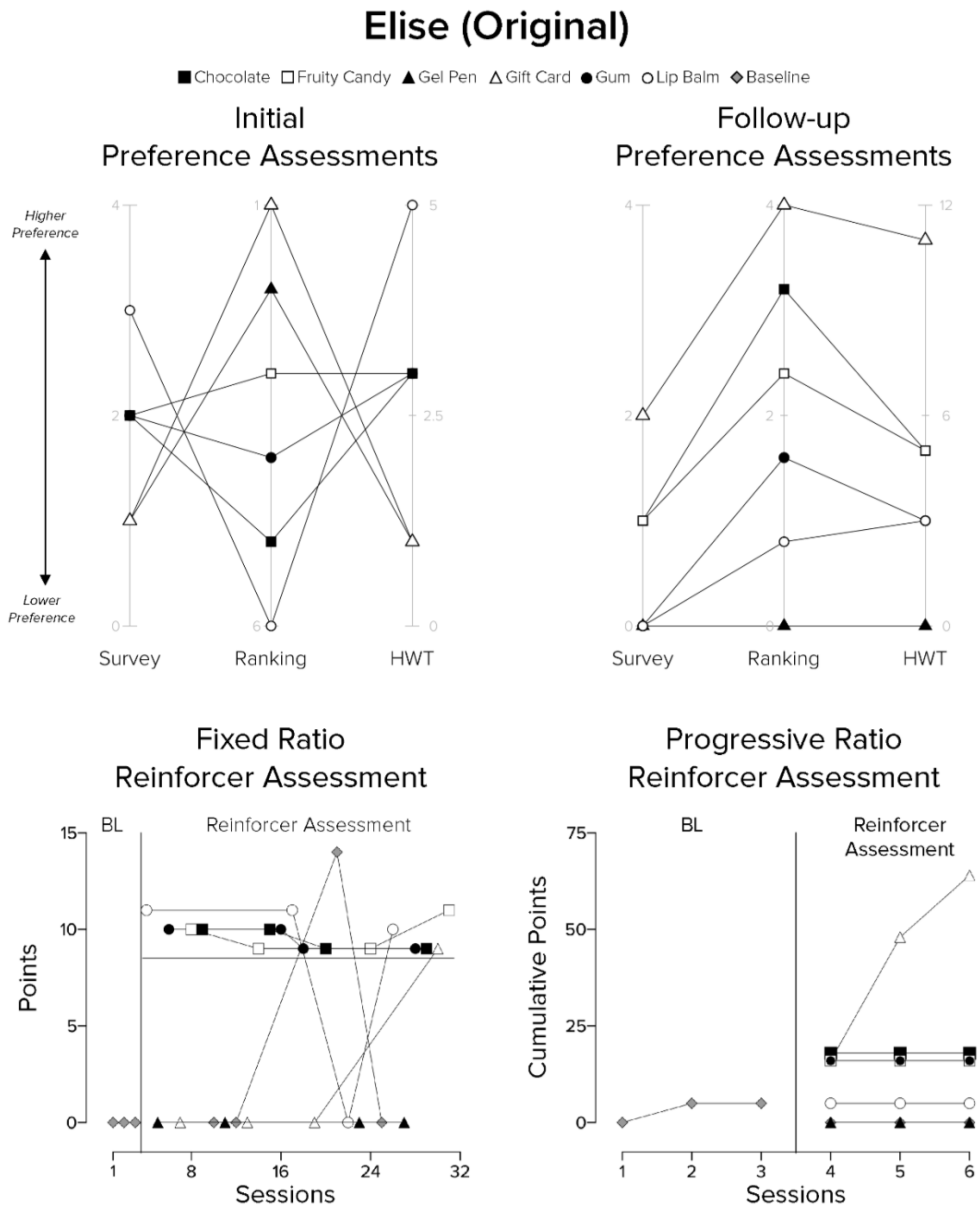
Assessment Data: Cindy



Note. HWT = hypothetical work task break point; BL = baseline

Figure 9.

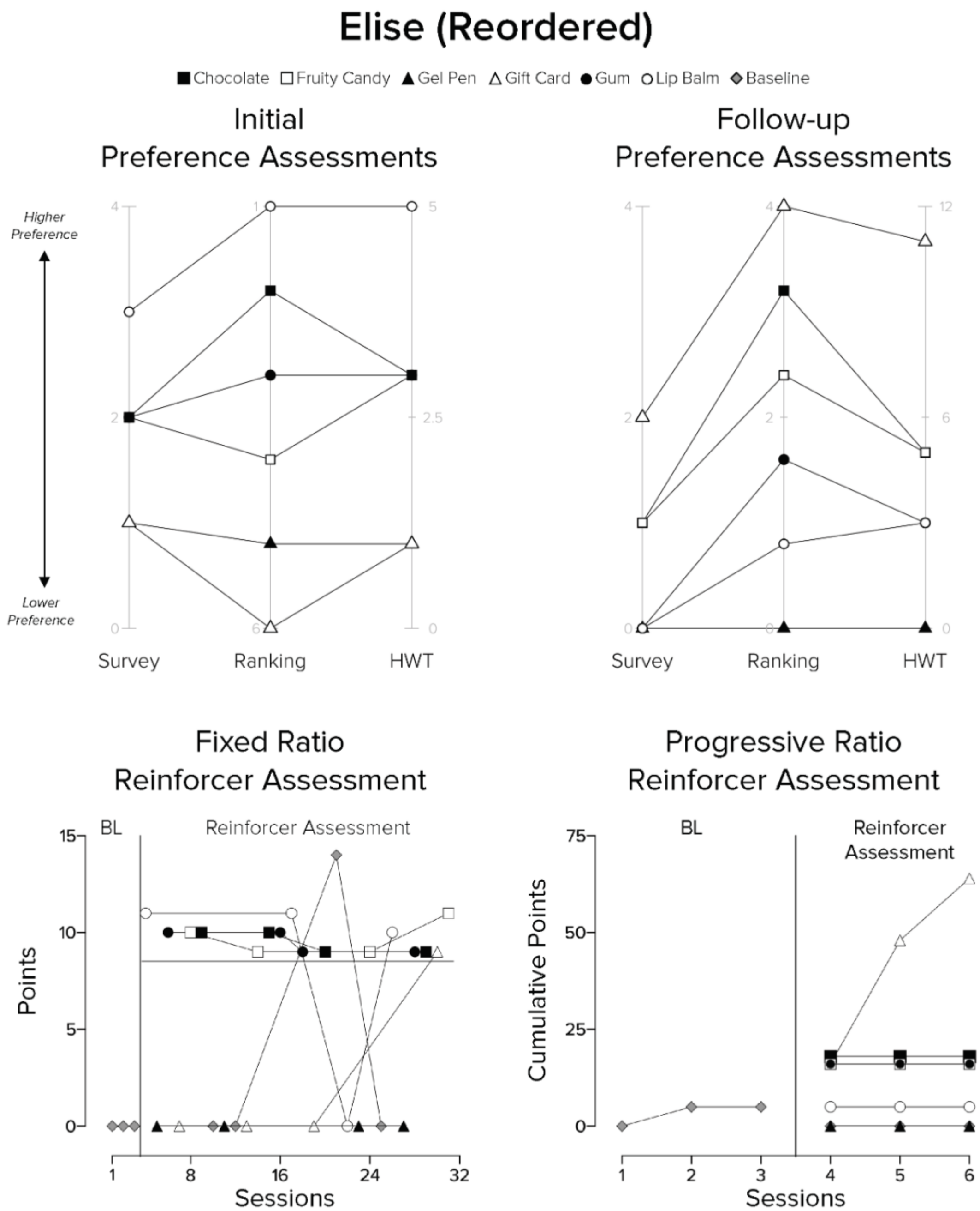
Assessment Data: Elise (Original)



Note. HWT = hypothetical work task break point; BL = baseline

Figure 10.

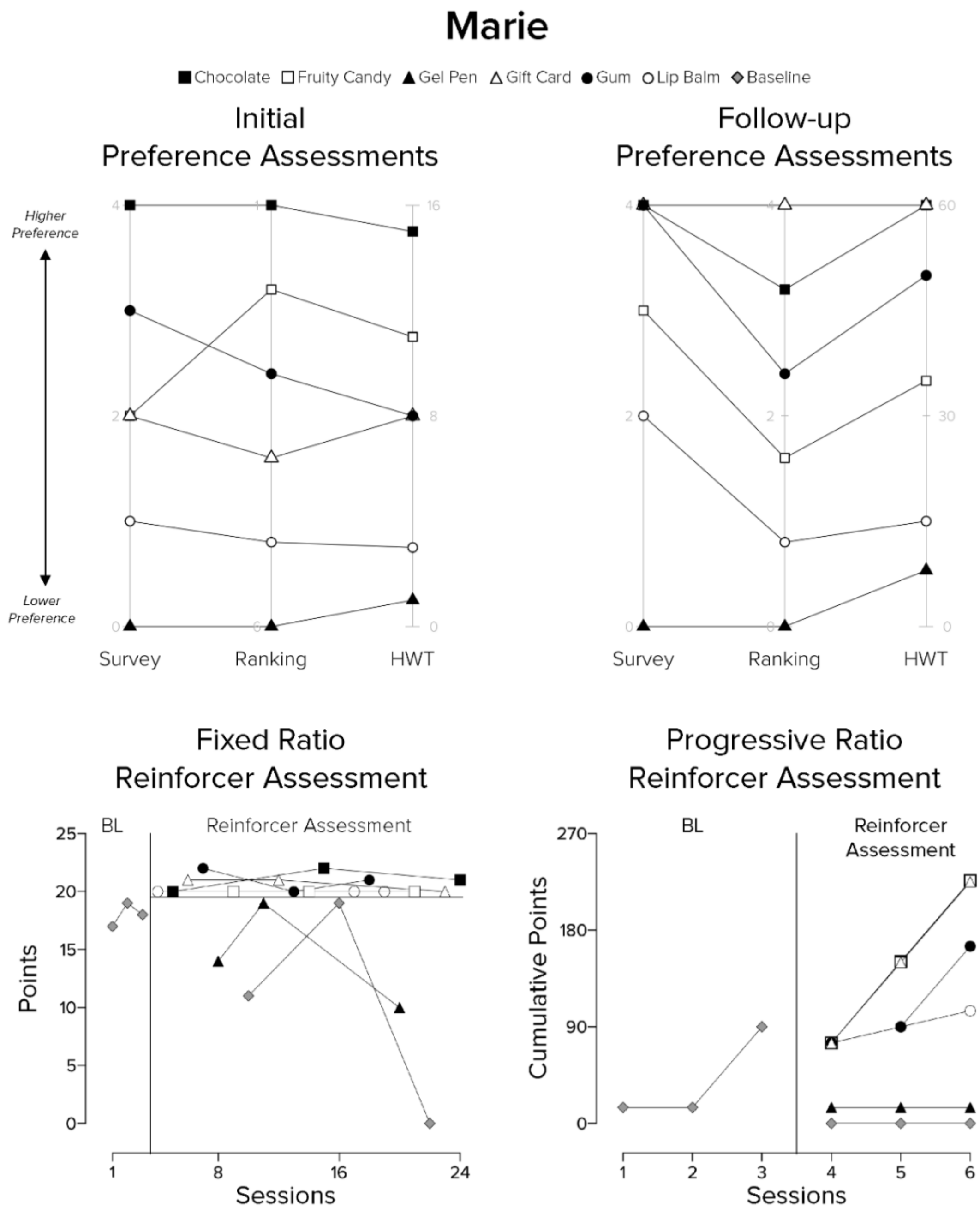
Assessment Data: Elise (Reordered)



Note. HWT = hypothetical work task break point; BL = baseline

Figure 11.

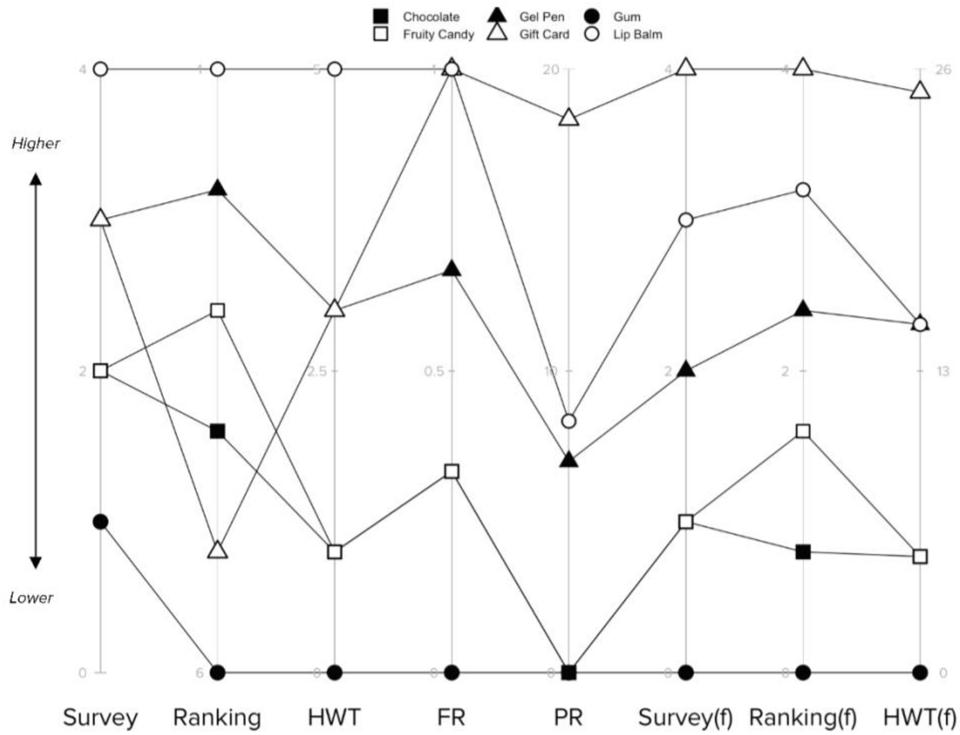
Assessment Data: Marie



Note. HWT = hypothetical work task break point; BL = baseline

Figure 12.

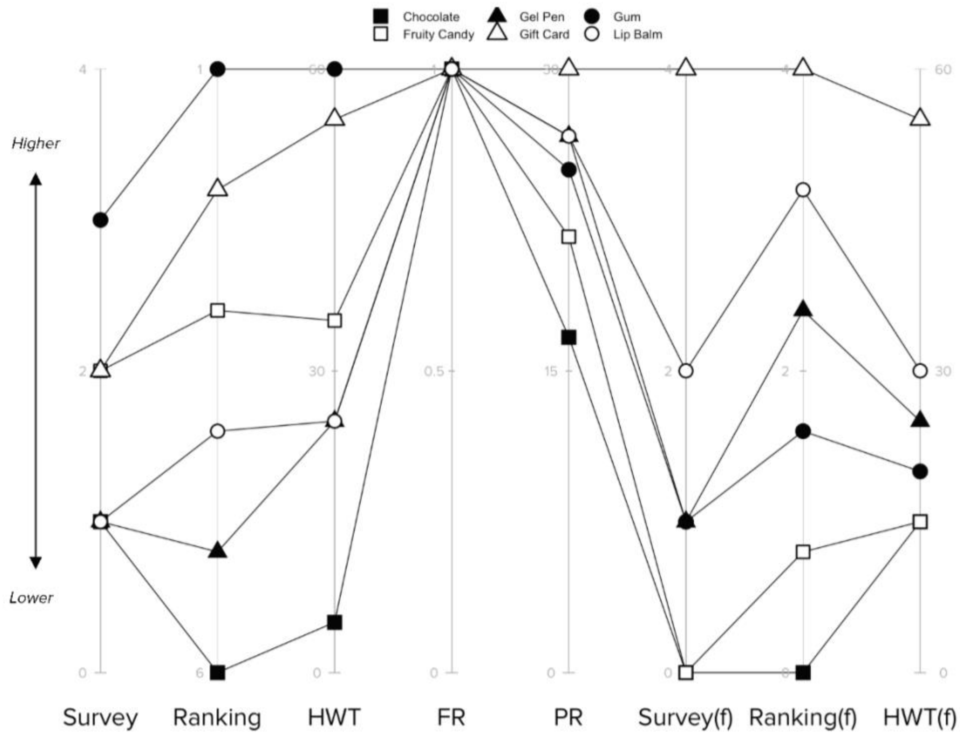
Outcome Summary: Julianne



Note. Measures toward the top of the scale indicate higher preference (for preference assessments), efficacy (for the fixed ratio reinforcer assessment), or effectiveness (for the progressive ratio reinforcer assessment). HWT = hypothetical work task break point; FR = fixed ratio reinforcer assessment (proportion of sessions incentive was earned); PR = progressive ratio break point; (f) = follow-up assessment.

Figure 13.

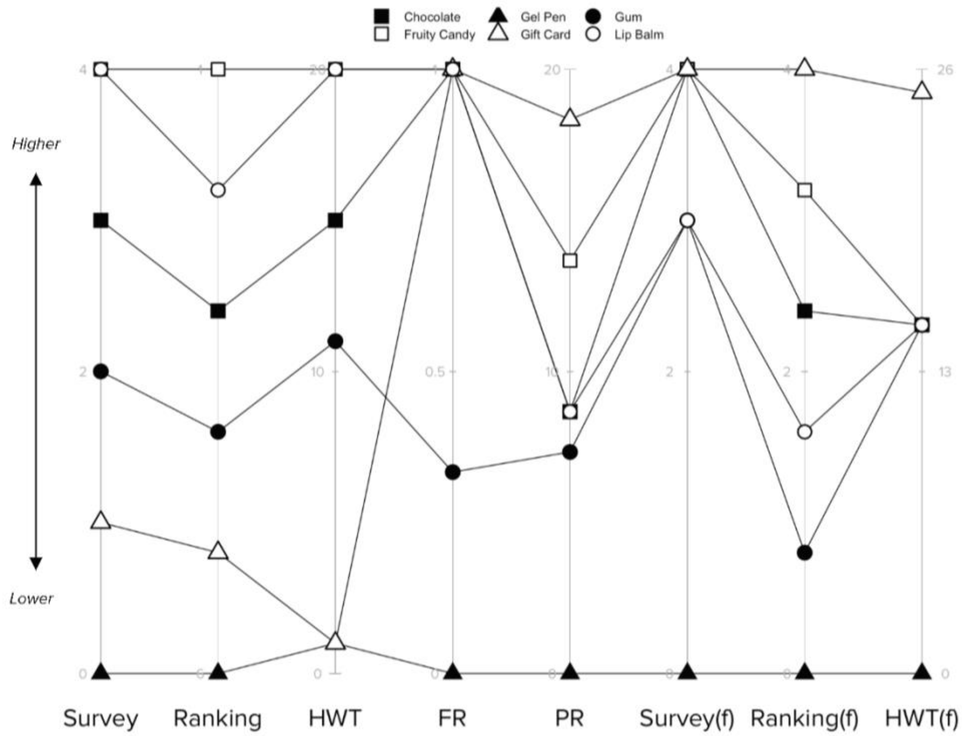
Outcome Summary: Shannon



Note. Measures toward the top of the scale indicate higher preference (for preference assessments), efficacy (for the fixed ratio reinforcer assessment), or effectiveness (for the progressive ratio reinforcer assessment). HWT = hypothetical work task break point; FR = fixed ratio reinforcer assessment (proportion of sessions incentive was earned); PR = progressive ratio break point; (f) = follow-up assessment.

Figure 14.

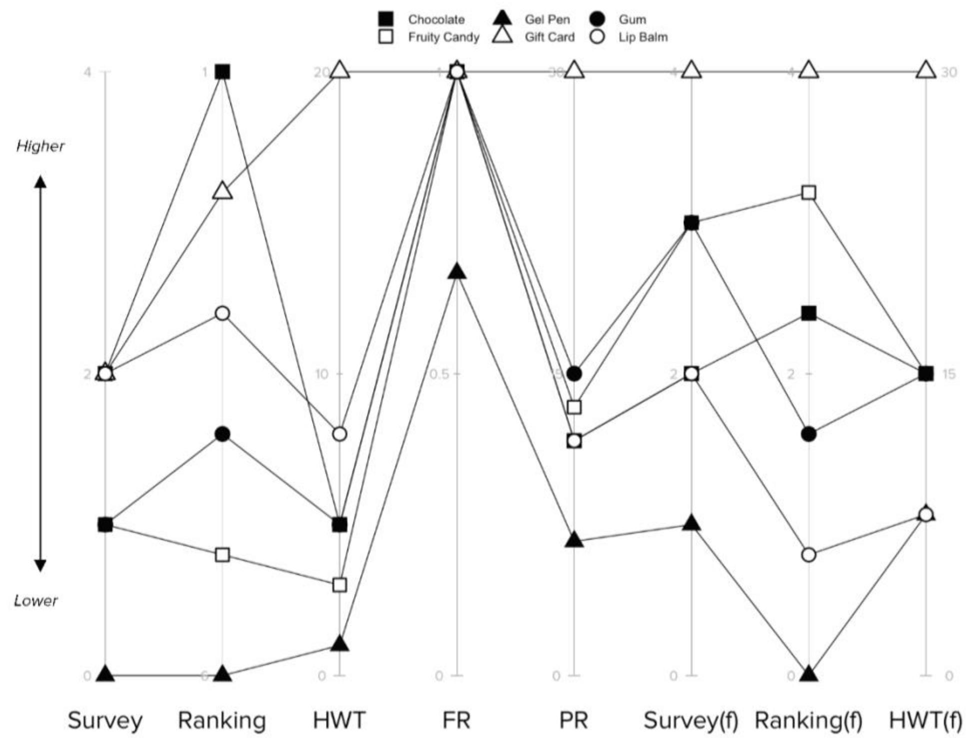
Outcome Summary: Alexia



Note. Measures toward the top of the scale indicate higher preference (for preference assessments), efficacy (for the fixed ratio reinforcer assessment), or effectiveness (for the progressive ratio reinforcer assessment). HWT = hypothetical work task break point; FR = fixed ratio reinforcer assessment (proportion of sessions incentive was earned); PR = progressive ratio break point; (f) = follow-up assessment.

Figure 15.

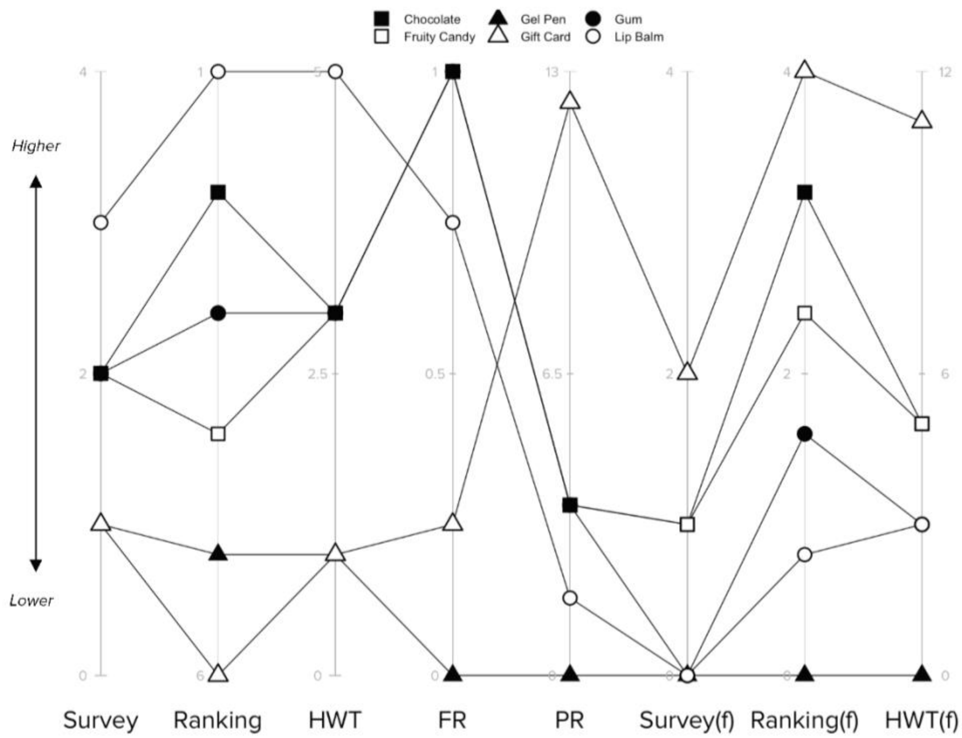
Outcome Summary: Cindy



Note. Measures toward the top of the scale indicate higher preference (for preference assessments), efficacy (for the fixed ratio reinforcer assessment), or effectiveness (for the progressive ratio reinforcer assessment). HWT = hypothetical work task break point; FR = fixed ratio reinforcer assessment (proportion of sessions incentive was earned); PR = progressive ratio break point; (f) = follow-up assessment.

Figure 16.

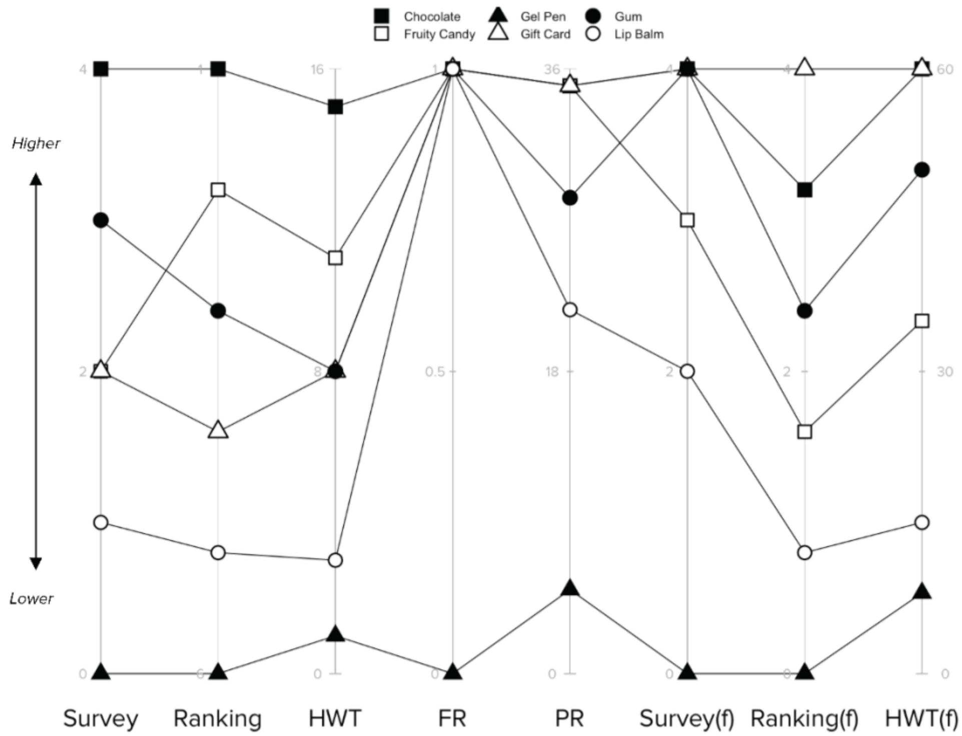
Outcome Summary: Elise (Reordered)



Note. Measures toward the top of the scale indicate higher preference (for preference assessments), efficacy (for the fixed ratio reinforcer assessment), or effectiveness (for the progressive ratio reinforcer assessment). HWT = hypothetical work task break point; FR = fixed ratio reinforcer assessment (proportion of sessions incentive was earned); PR = progressive ratio break point; (f) = follow-up assessment.

Figure 17.

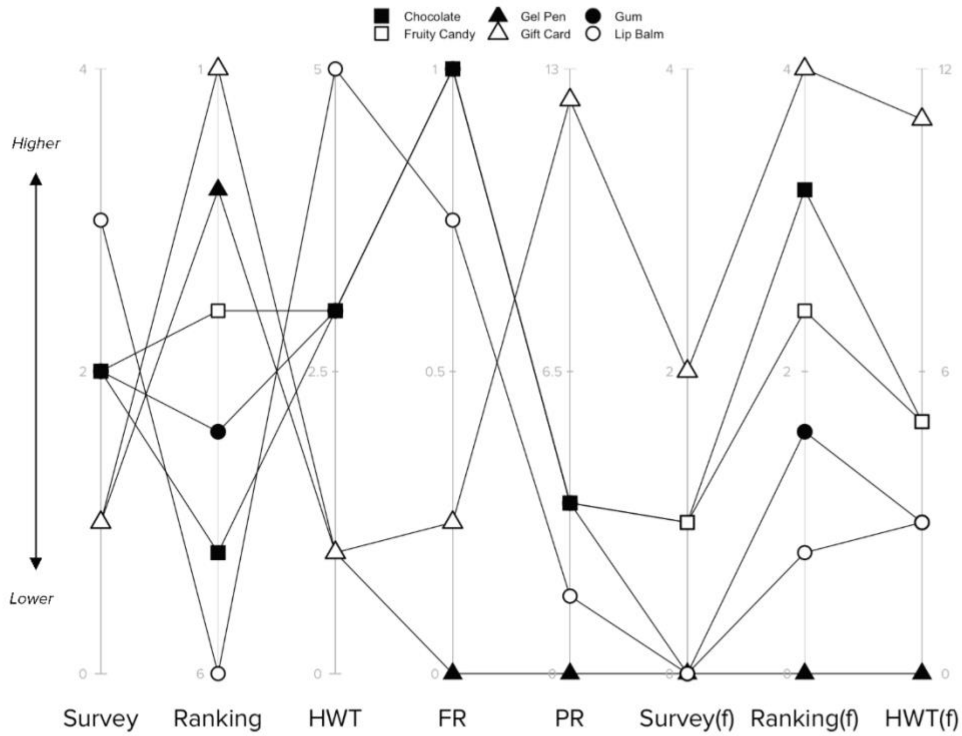
Outcome Summary: Marie



Note. Measures toward the top of the scale indicate higher preference (for preference assessments), efficacy (for the fixed ratio reinforcer assessment), or effectiveness (for the progressive ratio reinforcer assessment). HWT = hypothetical work task break point; FR = fixed ratio reinforcer assessment (proportion of sessions incentive was earned); PR = progressive ratio break point; (f) = follow-up assessment.

Figure 18.

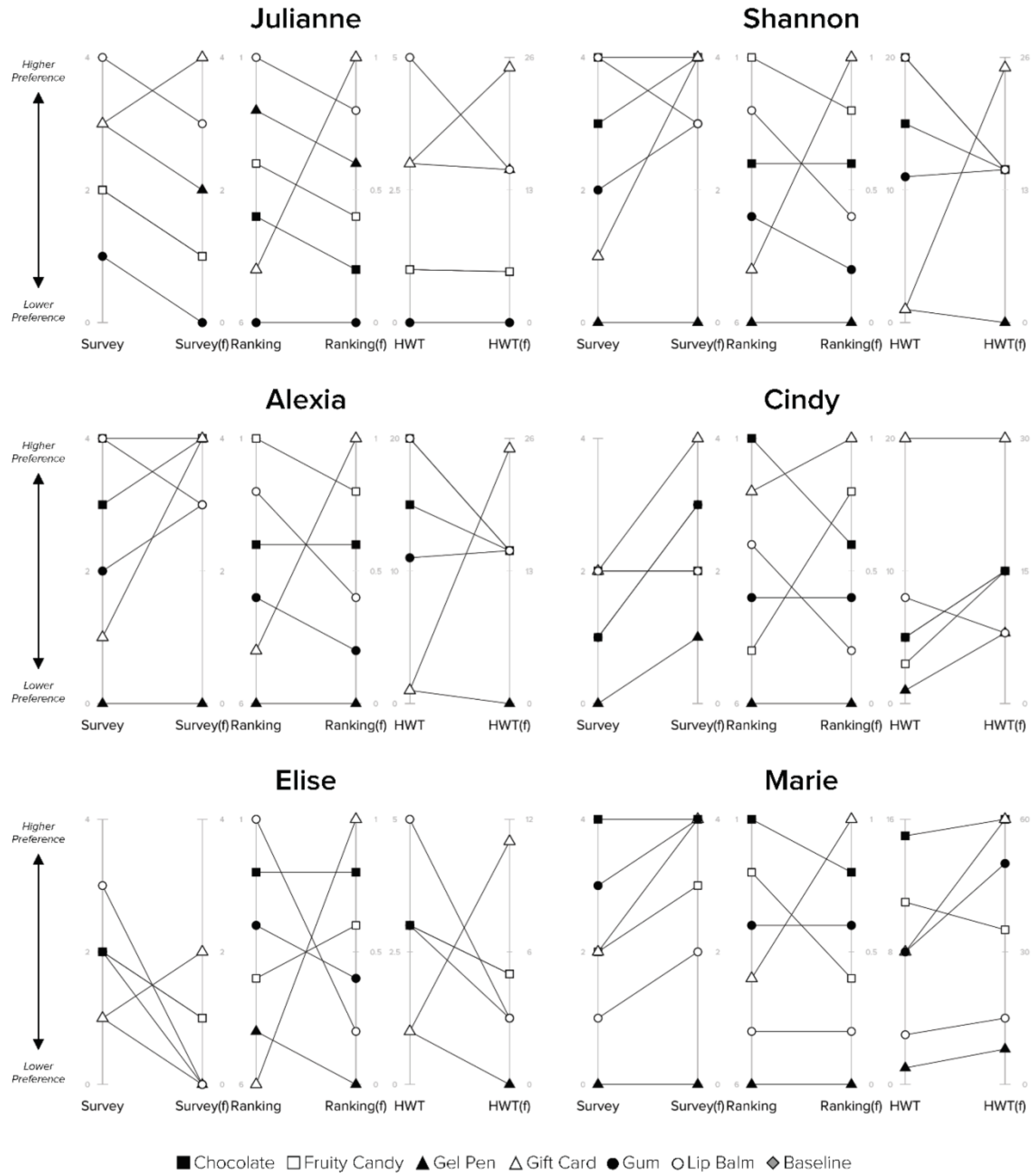
Outcome Summary: Elise (Original)



Note. Measures toward the top of the scale indicate higher preference (for preference assessments), efficacy (for the fixed ratio reinforcer assessment), or effectiveness (for the progressive ratio reinforcer assessment). HWT = hypothetical work task break point; FR = fixed ratio reinforcer assessment (proportion of sessions incentive was earned); PR = progressive ratio break point; (f) = follow-up assessment.

Figure 19.

Preference Assessment Test–Retest Data



Note. For each pairing, initial preference assessments are displayed on the left; follow-up preference assessments are displayed on the right. HWT = hypothetical work task break point; (f) = follow-up assessment.

Appendix A

Incentives Used in Reviewed Studies

Waldvogel & Dixon, 2008

1. $5 McDonalds gift card
2. $5 movie rental gift card
3. $5 phone card
4. $5 Starbucks gift card
5. $5 Wal-Mart gift card
6. 34-min break
7. Parking spot
8. Public recognition
9. Candy
10. T-shirt

Wilder et al., 2011

1. Cash
2. Gift card or tickets
3. Overnight hotel-stay
4. Free or discounted clothing
5. Day cruise
6. Spa treatment
7. Break or leave early
8. Praise from supervisor
9. Employee of the year or month
10. Employee games
11. Paid day off
12. More hours on schedule
13. Parking spot
14. Reduced workload
15. Scholarship nomination
16. Letter of recommendation
17. 25% off any item
18. BOGO any item
19. Lunch voucher
20. Luncheon or party
21. Assistant for a task
22. Coffee
23. Dessert
24. Snacks

*Incentives varied across participants.

Wilder et al., 2007

1. Cash
2. Best Buy gift card
3. Blockbuster gift card
4. Bookstore gift card
5. Bowling gift card
6. Home store gift card
7. Restaurant gift card
8. Shoe store gift card
9. Spa gift card
10. Surf shop gift card
11. Target gift card
12. Theme park gift card
13. Video rental gift card
14. Wal-Mart gift card
15. Baseball tickets
16. Movie passes
17. Theater tickets
18. Afternoon off of work
19. Floating holiday
20. Lunch out
21. Money for unit's use
22. Overtime hours
23. Paid time off
24. Travel to industry event
25. Industry recognition
26. Monthly recognition
27. Workplace recognition
28. Calligraphy supplies
29. Candy
30. Computer supplies
31. Cookies
32. Hair accessories
33. Lip gloss
34. Lottery tickets
35. Lunch voucher
36. Nail polish
37. Snacks
38. Soda or drinks

*Incentives varied across participants.

Wilder et al., 2006

1. $5 general store gift card
2. $5 movie theater gift card
3. Leave work 10 min early
4. Assistant for a task
5. Casual clothes
6. Snacks

Wine & Axelrod, 2014

1. $5 coffee gift card
2. $5 gas card
3. $5 music gift card
4. $5 restaurant gift card
5. Candy
6. Carbonated beverage
7. Lottery tickets
8. Office supplies

Wine et al., 2012

1. $5 convenience store gift card
2. $5 electronics gift card
3. $5 movie gift card
4. 10-min break
5. Skip post-shift meeting
6. Choose assigned work duties
7. Choose assigned work location
8. Take client on community outing
9. Candy

Wine et al., 2013, Study 1

1. $10 cash
2. 20-min break
3. Skip pre-shift meeting
4. Leave work 40 min early
5. Choose assigned work duties
6. Choose assigned work location
7. Parking space
8. Take client on community outing
9. Candy

Wine et al., 2013, Study 2

1. $10 cash
2. $10 convenience store gift card
3. $10 donut gift card
4. $10 electronics gift card
5. $10 fast food gift card
6. $10 movie rental gift card
7. $10 music gift card
8. Candy
9. Lottery tickets

Wine, Kelley, & Wilder, 2014

1. $5 coffee gift card
2. $5 movie theater gift card
3. $5 online store gift card
4. Campus dining pass
5. Coffee mug
6. Lottery tickets
7. On-campus entertainment tickets
8. Universal serial bus memory card

Wine, Reis, & Hantula, 2014, Experiments 1–2

1. $5 coffee gift card
2. $5 convenience store gift card
3. $5 movie rental gift card
4. $5 restaurant gift card
5. Candy
6. Lottery tickets
7. Office supplies

Wine & Wilder, 2009

1. $5 general store gift card
2. $5 health food gift card
3. $5 movie theater gift card
4. Specialty pencils
5. Specialty candy
6. Stationery

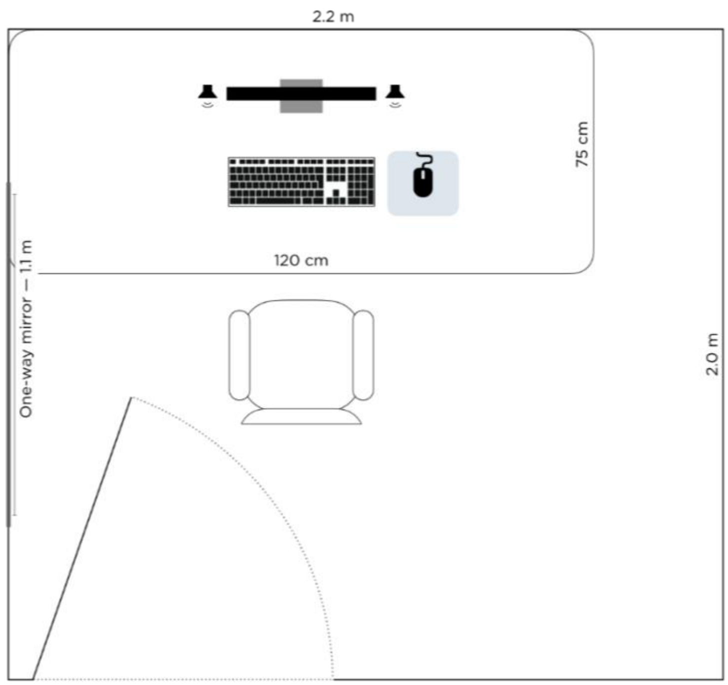
Appendix B

Participant Recruitment Script

"The purpose of the study is to evaluate the effects of several incentive systems on performance of a task. Participants will be asked to complete a basic computer task for about an hour per visit to the lab. During the course of the study, participants will have the opportunity to work for a variety of incentives. Participants will need to be available to return to the lab up to 10 times, for a total time commitment of about 10 hours. In exchange for participation, participants will receive extra credit points equal to 0.5% of your final course grade for each hour of participation. Additionally, participants will any incentives they earn during the course of the study. In order to be eligible to participate, you must be 18 years of age or older."

Appendix C

Session Room Layout



Appendix D

Preference Assessment Stimuli

Name	Single/ Choice	Item (*size*)	Value
Chocolate	Choice	Snickers (*93.3 g*)	$1.00
		Twix (*50.7 g*)	$0.63
Fruit Candy	Choice	Skittles (*61.5 g*)	$0.56
		Starburst (*58.7 g*)	$0.67
Gel Pen	Single	Pilot G–2 Pen (*0.7 line-weight*)	$0.99
Gift Card	Single	Amazon Gift Card	$2.00
Gum	Choice	5 Peppermint Gum (*15 sticks*)	$0.97
		5 Spearmint Gum (*15 sticks*)	$0.97
Lip Balm	Choice	Burt's Bees Lip Balm (*4.25 g*)	$2.39
		Carmex Lip Balm (*10 g*)	$2.19

Note: Dollar values were determined by searching for each item on Amazon and calculating a unit price based off of the "Amazon's Choice" option for all incentives (except for the gift card).

Appendix E

Incentive-Stimuli Tokens

1 CHOCOLATE BAR

1 PACK OF FRUIT-FLAVORED CANDY

1 GEL INK PEN

1 $2 AMAZON GIFT CARD

1 PACK OF GUM

1 STICK/TUBE OF LIP BALM

Appendix F

Debrief Statements

Following administration of the follow-up preference assessments, the experimenter conducted an unstructured debrief, asking participants to provide context on why they worked for some items and did not work for others. Statements based off of experimenter notes and thus are not verbatim.

Julianne

Worked for:

- gift card, because she had no money
- lip balm, because she could give it to her friends
- gel pens, because they are her favorite type

Shannon

No statements recorded

Alexia

Worked for items based on what she already had

Cindy

Worked for:

- gift card, because she had a lot of stuff saved in her amazon card
- gel pens, because she wanted them for her job as a waitress (they are her favorite kind)
- gum, because it is her favorite kind

Elise

- Day 1 did not complete work because she had an essay due for another class
- Day 3 she completed work because she had nothing else to do
- Day 4 she had a lot of everything already but no gift cards, so she only worked for the gift card

Marie

- Was counting how many of each item she had and got to a point where she did not need any more gum
- Was kind of hungry on the last day, which probably played a role

Appendix G

Informed Consent Form

INFORMED CONSENT STATEMENT

An Evaluation of Preference Assessment Methods on Work Outcomes

INTRODUCTION

The Department of Applied Behavioral Science at the University of Kansas supports the practice of protection for human subjects participating in research. The following information is provided for you to decide whether you wish to participate in the present study. You may refuse to sign this form and not participate in this study. You should be aware that even if you agree to participate, you are free to withdraw at any time. If you do withdraw from this study, it will not affect your relationship with this unit, the services it may provide to you, or the University of Kansas.

PURPOSE OF THE STUDY

The purpose of the study is to evaluate different methods of assessing preference for various workplace incentives.

PROCEDURES

By participating in this study, you will be asked to complete a demographic survey. Following completion of this survey, you will sample a computer task (clicking on moving boxes) that you will use throughout the experiment. You will then complete several surveys asking you to indicate your preference for a number of items and activities. Next, you will be given the opportunity to complete the basic computer task, which will be presented for several brief trials. At the beginning of each trial, the experimenter will inform you of the nature of the incentive system in place for the trial. Any incentives you earn will be delivered as a voucher at the end of each trial; vouchers will be traded in for the specific items earned at the end of each 60-minute experimental session. You will be asked to return to the lab up to 10 times for a total time commitment of no more than 10 hours. All sessions may be recorded. Video recordings will be used by the research team solely for the purposes of collecting data and the recordings will not be made available to anyone other than the research team. The files will be kept securely in electronic format on a secure server to which only the research team has access and will be destroyed after a maximum of five years.

RISKS

There are no risks associated with participation in this study.

BENEFITS

Beyond compensation, there are no direct benefits to participants. The present study will contribute to the literature base on characteristics of effective incentive systems and will benefit managers in organizational settings by informing practice in effective performance management.

Rev 10/17

PAYMENT TO PARTICIPANTS

You will be compensated by receiving 0.5% points of extra credit for every hour of participation, which will be applied to your undergraduate ABSC course from which you were recruited.

PARTICIPANT CONFIDENTIALITY

Your name will not be associated in any publication or presentation with the information collected about you or with the research findings from this study. Instead, the researcher(s) will use a study number or a pseudonym rather than your name. Your identifiable information will not be shared unless (a) it is required by law or university policy, or (b) you give written permission.

Permission granted on this date to use and disclose your de-identified information remains in effect indefinitely. By signing this form, you give permission for the use and disclosure of your de-identified information for the purposes of this study at any time in the future.

REFUSAL TO SIGN CONSENT AND AUTHORIZATION

You are not required to sign this Consent and Authorization form and you may refuse to do so without affecting your right to any services you are receiving or may receive from the University of Kansas or to participate in any programs or events of the University of Kansas. However, if you refuse to sign, you cannot participate in this study.

CANCELLING THIS CONSENT AND AUTHORIZATION

You may withdraw your consent to participate in this study at any time. You also have the right to cancel your permission to use and disclose further information collected about you, in writing, at any time, by sending your written request to: *Florence D. DiGennaro Reed, 4020 Dole, 1000 Sunnyside Avenue, Lawrence, KS 66045.*

If you cancel permission to use your information, the researcher(s) will stop collecting additional information about you. However, the research team may use and disclose information that was gathered before they received your cancellation, as described above.

QUESTIONS ABOUT PARTICIPATION

Questions about procedures should be directed to the researchers listed at the end of this consent form.

PARTICIPANT CERTIFICATION:

I have read this Consent and Authorization form. I have had the opportunity to ask, and I have received answers to, any questions I had regarding the study. I understand that if I have any additional questions about my rights as a research participant, I may call (785) 864-7429 or (785) 864-7385, write the Human Research Protection Program (HRPP), University of Kansas, 2385 Irving Hill Road, Lawrence, Kansas 66045-7568, or email irb@ku.edu.

I agree to take part in this study as a research participant. By my signature, I affirm that I am at least 18 years old and that I have received a copy of this Consent and Authorization form.

_______________________________ _____________________
Print Participant's Name Date

____
Participant's Signature

RESEARCHER CONTACT INFORMATION

Matthew D. Novak
Primary Investigator
4085 Dole Human Development Center
University of Kansas
1000 Sunnyside Avenue
Lawrence, KS 66045
novak@ku.edu

Florence D. DiGennaro Reed, Ph.D., BCBA-D
Faculty Supervisor
4020 Dole Human Development Center
University of Kansas
1000 Sunyside Avenue
Lawrence, KS 66045
fdreed@k.edu
785.864.0521

Appendix H

Demographic Questionnaire

Demographic Questionnaire

ID # (given to you by researcher):
Do NOT enter your student ID number!

Age:______________

Gender with which you identify:
- ◯ Female
- ◯ Male
- ◯ Other: ____________________

Race/ethnic background:
- ◯ White/Caucasian
- ◯ Black/African American
- ◯ Hispanic/Latino
- ◯ Asian
- ◯ Native American
- ◯ Pacific Islander
- ◯ Mixed
- ◯ Other

Do you have a DOCUMENTED disability?
- ◯ Yes
- ◯ No

If yes, please specify:
- ◯ Learning disability
- ◯ Attention deficit/hyperactivity disorder
- ◯ Physical disability
- ◯ Other

Is English your first/primary language?
- ◯ Yes
- ◯ No

In what ways do you earn income? Mark all that apply.
- ◯ Paid employment
- ◯ Family members (parents, partners, etc)
- ◯ Student loans
- ◯ Other: ______________________________
- ◯ None

Appendix I

Fixed Ratio Session Orders

Session	Phase	Julianne	Shannon	Alexia	Cindy	Elise	Marie
1	BL	BL	BL	BL	BL	BL	BL
2	BL	BL	BL	BL	BL	BL	BL
3	BL	BL	BL	BL	BL	BL	BL
4	R	Gel pen	Choc.	Choc.	Lip balm	Lip balm	Lip balm
5	R	GC	Fruit	Fruit	Choc.	Gel pen	Choc.
6	R	Choc.	Gum	Gel pen	Gum	Gum	GC
7	R	Fruit	GC	Lip balm	GC	GC	Gum
8	R	Gum	Lip balm	Gum	Gel pen	Fruit	Gel pen
9	R	Lip balm	Gel pen	GC	Fruit	Choc.	Fruit
10	R	BL	BL	BL	BL	BL	BL
11	R	Fruit	GC	GC	Gum	Gel pen	Gel pen
12	R	GC	BL	Gum	BL	BL	GC
13	R	Gum	Lip balm	BL	Fruit	GC	Gum
14	R	Choc.	Fruit	Lip balm	GC	Fruit	Fruit
15	R	Gel pen	Choc.	Gel pen	Lip balm	Choc.	Choc.
16	R	BL	Gel pen	Choc.	Gel pen	Gum	BL
17	R	Lip balm	Gum	Fruit	Choc.	Lip balm	Lip balm
18	R	BL	Gel pen	Lip balm	Lip balm	Gum	Gum
19	R	Gum	Gum	Fruit	Fruit	GC	Lip balm
20	R	Gel pen	Lip balm	Gum	Gel pen	Choc.	Gel pen
21	R	Lip balm	BL	BL	BL	BL	Fruit
22	R	Choc.	Fruit	Gel pen	GC	Lip balm	BL
23	R	GC	GC	Choc.	Choc.	Gel pen	GC
24	R	Fruit	Choc.	GC	Gum	Fruit	Choc.
25	R					BL	
26	R					Lip balm	
27	R					Gel pen	
28	R					Gum	
29	R					Choc.	
30	R					GC	
31	R					Fruit	

Note: BL = Baseline; R = Reinforcer assessment; Choc. = Chocolate; Fruit = Fruity candy; GC = Gift card.

Appendix J

Progressive Ratio Session Orders

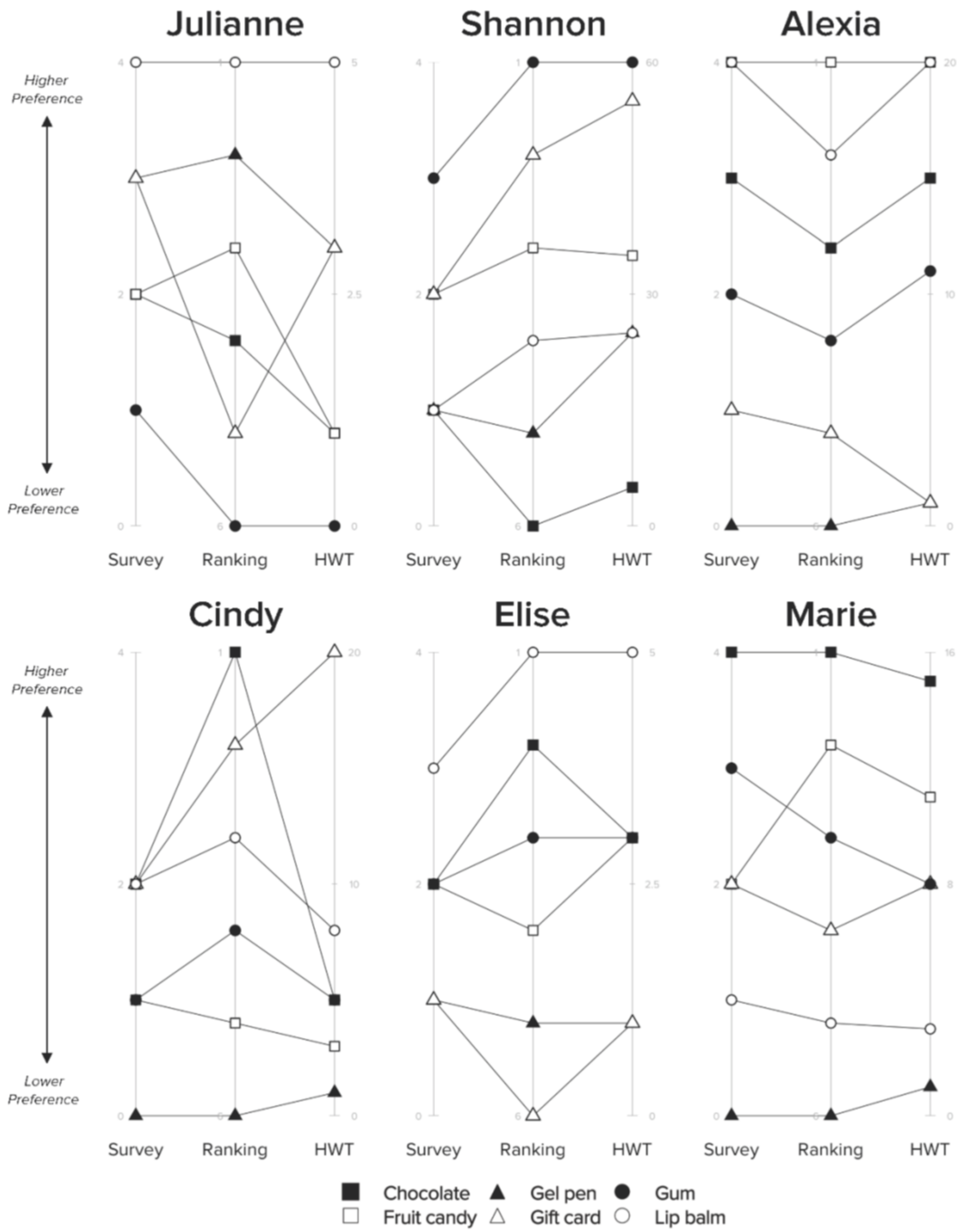
Session	Phase	Julianne	Shannon	Alexia	Cindy	Elise	Marie
1	BL	BL	BL	BL	BL	BL	BL
2	BL	BL	BL	BL	BL	BL	BL
3	BL	BL	BL	BL	BL	BL	BL
4	R	Fruit	Fruit	Fruit	Gel pen	Gel pen	Lip balm
5	R	Gum	Gel pen	Gum	Choc.	Gum	Choc.
6	R	GC	GC	Choc.	Fruit	Fruit	Fruit
7	R	Lip balm	Lip balm	GC	Gum	Lip balm	Gel pen
8	R	Choc.	Gum	Lip balm	Lip balm	Choc.	GC
9	R	Gel pen	*Choc.	Gel pen	GC	GC	Gum
10	R	BL	BL	BL	BL	BL	BL
11	R	GC	GC	Fruit	Choc.	Gum	Choc.
12	R	BL	BL	BL	Fruit	BL	Gum
13	R	Gum	Fruit	Lip balm	GC	GC	Fruit
14	R	Gel pen	Gum	Gel pen	Lip balm	Fruit	Gel pen
15	R	Fruit	Lip balm	Choc.	Gel pen	Choc.	GC
16	R	Choc.	Choc.	GC	Gum	Lip balm	Lip balm
17	R	Lip balm	Gel pen	Gum	BL	Gel pen	BL
18	R	Gum	GC	Gum	Fruit	Lip balm	Lip balm
19	R	Choc.	Gel pen	BL	Gel pen	Choc.	Gum
20	R	Fruit	BL	Lip balm	GC	Gel pen	Choc.
21	R	GC	Choc.	Fruit	BL	GC	Gel pen
22	R	BL	Lip balm	Gel pen	Choc.	Fruit	GC
23	R	Lip balm	Gum	Choc.	Gum	BL	BL
24	R	Gel pen	Fruit	GC	Lip balm	Gum	Fruit

Note: BL = Baseline; R = Reinforcer assessment; Choc. = Chocolate; Fruit = Fruity candy; GC = Gift card.

* Procedural error occurred prior to this session (completed one chocolate session with incorrect progressive ratio schedule).

Appendix K

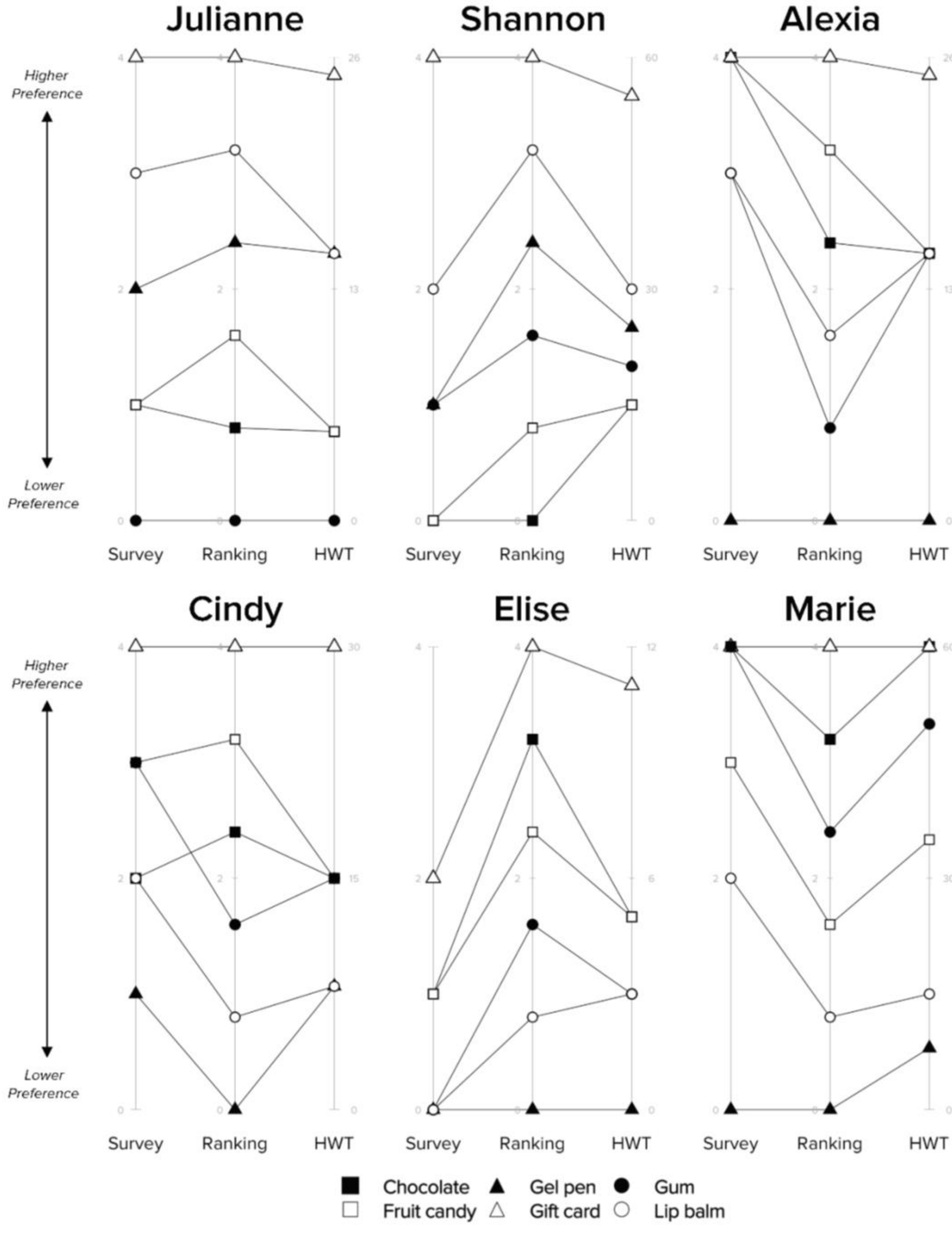
Initial Preference Assessment Data



Note. For Elise, data displayed are reordered rankings. HWT = hypothetical work task break point.

Appendix L

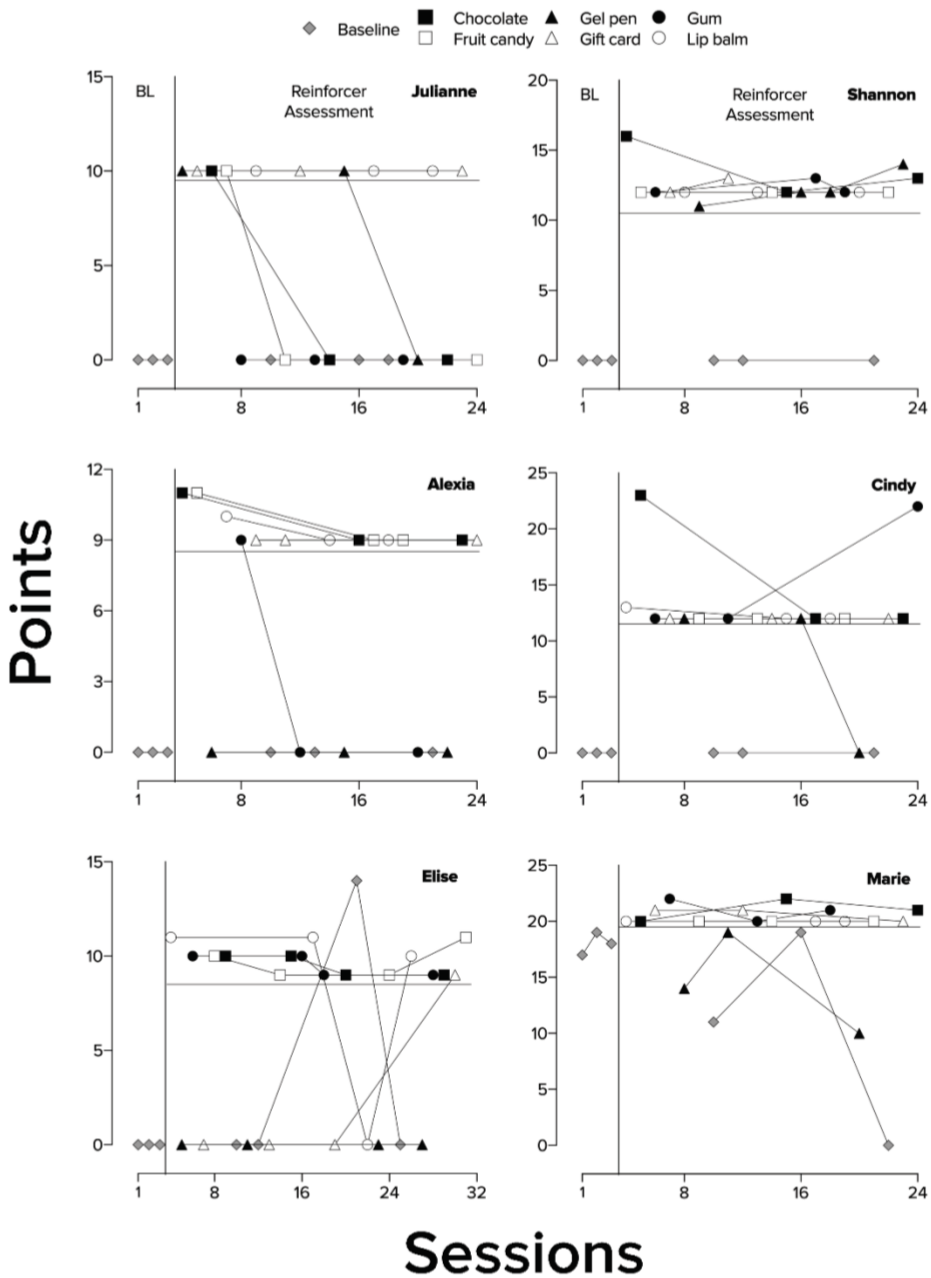
Follow-up Preference Assessment Data



Note. HWT = hypothetical work task break point.

Appendix M

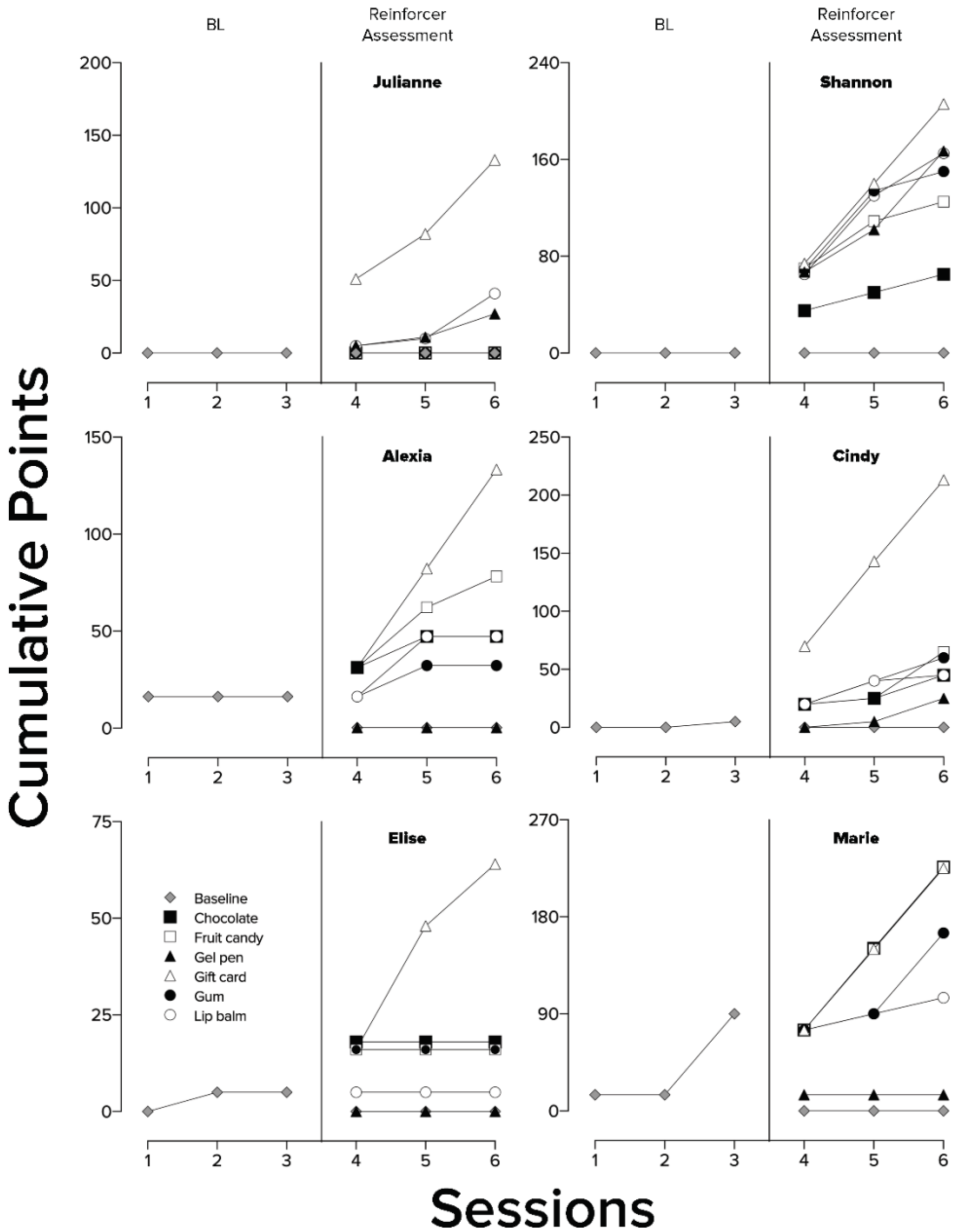
Fixed Ratio Assessment Data



Note: BL = baseline.

Appendix N

Progressive Ratio Assessment Data



Note: BL = baseline.