ABSTRACT

THE PREDICTIVE VALIDITY OF STIMULUS PREFERENCE ASSESSMENTS

By Sarah A. Stephan

This study examined the predictability and stability of self-nomination and verbal forced choice reinforcer assessment methods. Reinforcers identified by self-nomination and verbal forced choice were compared to experimentally-established reinforcers to determine the ability of these two reinforcer assessment methods to accurately predict effective reinforcers. The study used an academically relevant maze passage to determine actual reinforcers. These methods were repeated four times within subjects to determine stability over time. These methods were completed with 15 elementary-age children in an analogue classroom setting. The results were analyzed using Kappa to eliminate chance agreement. Verbal forced choice was found to have more significant predictability than the nomination survey. When the procedures were replicated multiple times, nomination survey became more accurate, but verbal forced choice became more accurate as well. Ultimately, verbal forced choice was the most effective method for predicting reinforcers.
THE PREDICTIVE VALIDITY OF STIMULUS PREFERENCE ASSESSMENTS

A Thesis

Submitted to the Faculty of Miami University
in partial fulfillment of
the requirements for the degree of
Educational Specialist
Department of Educational Psychology
by
Sarah A. Stephan
Miami University
Oxford, Ohio
2008

Advisor______________________________________
Dr. Kevin M. Jones

Reader_______________________________________
Dr. Katherine Wickstrom

Reader_______________________________________
Dr. Steuart Watson

Reader_______________________________________
Dr. Tonya Watson
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Method</td>
<td>7</td>
</tr>
<tr>
<td>Results</td>
<td>12</td>
</tr>
<tr>
<td>Discussion</td>
<td>15</td>
</tr>
<tr>
<td>References</td>
<td>20</td>
</tr>
<tr>
<td>Tables</td>
<td>23</td>
</tr>
<tr>
<td>Figure</td>
<td>26</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

Table 1. Descriptive data and reinforcer assessment results for each participant……23
Table 2. Accuracy and kappa coefficients for nomination and verbal forced choice....24
Table 3. Frequency of classification rates and improvement across four trials for the two preference assessment procedures…………………………………………………..25
LIST OF FIGURES

Figure 1. An example of three types of response patterns observed during the reinforcer assessment…………………………………………………………………………………………..26
DEDICATION/ACKNOWLEDGEMENT

I would like to sincerely thank everyone who assisted with the collection of data for this project: Chelsey Schneider, Brianna Sarr, Patricia Rodriguez, Mike Albrecht, Nathan Dumford, Erin Vincent, and Laura Russ, as well as the Miami University 2007 school psychology cohort. This project would not exist without the excellent guidance and support of Dr. Kevin M. Jones. I would like to thank my thesis committee, including Dr. Katherine Wickstrom, Dr. Steuart Watson, and Dr. Tonya Watson for offering their sage advice. Additionally, I would like to thank my parents, Stan and Claudia Stephan, for their unyielding confidence in me.

I would like to dedicate this thesis to the students of the School of R.O.C.K., who participated in this study without complaint, even when they did not want to work for the control coupon.
Introduction

Contingency management is one of the most effective and empirically supported treatment paradigms for children and adolescents. Meta-analyses indicate that reinforcement-based treatments are among the most powerful approaches to increasing academic engagement (Weaver, 2004), decreasing disruptive classroom behavior (Stage & Quiroz, 1997), managing ADHD (DuPaul & Eckert, 1997), and specializing instruction (Lloyd, Forness, & Kavale, 1998).

These findings would suggest that it is relatively easy to design behavior management programs in the classroom – simply provide reinforcement contingent on appropriate, alternative behaviors. The challenge, however, is identifying those stimuli, events, or things that will actually function as reinforcers for a particular child, at a particular point in time. This is a challenge because events that a person prefers or finds appealing may have very little impact on their actual behavior. A child, for example, may report that he or she “likes” to play Game Boy, but access to a new game card or more playing time after completing a chore does not actually increase the rate of responding. Also, the effectiveness of an actual reinforcer may vary according to establishing operations (e.g., immediacy, size, satiation/deprivation) and response effort required. For example, contingent access to dessert might sufficiently sustain a child’s eating of vegetables, but it may be less effective when the dessert is small, when dessert time interferes with play time, or when the vegetable is broccoli or squash.

The idiosyncratic effects of reinforcers are a serious challenge for educators who have instructional routines that allow little time for identifying, dispensing, and substituting reinforcers. For practical reasons, teachers may prefer tangible rewards such as awards or prizes that can be dispensed quickly or, alternatively, reward tokens or coupons that can be distributed quickly but are “cashed in” at a later time or in a different setting (e.g., home). Rewards that require disruptions in routines, such as reduced work, or require teacher supervision, such as cooperative games, may be equally effective but much more intrusive.

Despite these challenges, the appropriate use of extrinsic rewards is an essential element of effective classroom management (Maag, 1999). From the teacher’s perspective, convenient methods for identifying those stimuli that actually function as
positive reinforcement are often needed to sustain student engagement and skill acquisition. From the child’s perspective, the stakes are much higher: If positive reinforcement is weak or missing from a classroom management system, the teacher may rely more heavily on coercive methods such as punishment or negative reinforcement (Maag, 2001).

The purpose of this study was to examine two methods of reinforcer selection. The first method was a reinforcer nomination survey (“How much would you like to ______”), which required the child to rate the perceived reinforcement value of each item in a list. The survey method is intuitive, quick, and arguably the most common approach to reinforcer selection. The second method was a verbal forced choice survey (e.g., “Would you rather ______ or ______”), which required the child to choose, with each item being paired twice, in a different order, against all others on the list. While both approaches are an option for most children, recent work has suggested that the forced choice method is more accurate in terms of predicting which stimuli actually function as positive reinforcement during a subsequent experimental analysis (Northup, Jones, Broussard, & George, 1995; Northup, George, Jones, Broussard, & Vollmer, 1996). Unlike previous research, however, the current study evaluated the relative accuracy of each method across time and exposure to contingencies.

**Defining Rewards, Preferences, and Reinforcers**

In reviewing the literature on stimulus preference assessments, it is necessary to distinguish the terms “reward,” “preference,” and “reinforcer.” A reward is a term used most often to describe a stimulus that is presented following a behavior, with an implicit assumption that the stimulus will increase its frequency. Rewards are defined topographically, and thus may or may not actually increase the frequency of the target response. A preference is a stimulus identified through a choice paradigm, which might include choosing among stimuli (e.g., nominations), choosing a rating associated with a stimulus (e.g., survey), or choosing between stimuli (e.g., forced choice). Stimulus preferences, like rewards, may or may not function as reinforcers. A reinforcer is a stimulus that is presented following a behavior, and increases its frequency. Thus, a reinforcer is identified through an experimental analysis.

**Methods of Preference Assessments for Typically Developing Children**
The most common method of preference assessment among typically developing children is the nomination or survey approach. A primary example is Fantuzzo, Rohrbeck, Hightower, and Work (1991), who explored the use of reinforcers in schools. They wanted to examine how much rewards were used and the relationship between the rewards teachers use and the rewards students prefer. They used a nomination system wherein the students involved were asked whether they preferred a reward “not at all,” “a little,” or “a lot” for the 86 items that were identified by their teachers. They found that there is a high use of rewards in elementary schools, children preferred a wide variety of rewards with no significant difference due to grade level, and the most shocking result of all was that there was no significant relationship between rewards teachers use and rewards children want. This is significant because the rewards being used in classrooms today may not be the most effective reinforcers.

Forced-choice methods are less common, but have been validated more extensively, particularly their use with developmentally delayed individuals who have limited verbal or communication skills. This method was originally developed to improve the quality of life for individuals by providing greater access to preferred activities and events (Hagopian, Long, & Rush 2004). Subsequent research has evaluated the accuracy of different forced-choice methods in predicting which stimuli actually function as positive reinforcement during a subsequent reinforcer assessment.

For example, Didden and de Moor (2004) conducted a study utilizing a forced choice assessment with toddlers with developmental and physical disabilities and a control group of typically developing toddlers. Each child went through a pair-wise comparison, where he or she was presented with two potentially reinforcing items, all of which were toys. Whichever item they interacted with first was deemed to be the reinforcer. They repeated these sessions until each child had shown a preference between each pair of items and each item had been compared to all other items, thereby identifying a rank order of reinforcers. Results showed that children with disabilities showed much more discrepancy in the rank order of reinforcers, meaning that they demonstrated stronger preferences. Also, children with disabilities preferred dynamic toys more than the non-disabled population.
Until recently, there was little research investigating the accuracy of preference assessments for typically developing children who have adequate communication skills. Perhaps this was because educators and clinicians assumed that a child with verbal skills could simply state what is likely to be reinforcing for them, either by choosing among available stimuli or rating the reinforcement value of an item.

In one of the first studies of stimulus preference assessments using children without cognitive disabilities, Northup, Jones, Broussard, and George (1995) compared nomination, verbal forced choice, and direct observation methods for 10 children with ADHD. The nomination procedure consisted of presenting five toys, and asking the child to select a favorite. The verbal forced-choice method required the child to choose between pairs of rewards read verbally to child, with each contrast presented twice (in a different order). The “preferred” toy was determined as the one selected most often. The direct observation method involved a 10-min free play arrangement, in which all five toys were available, and the “preferred” toy was determined by ranking the percentage of time engaged with each toy. Next, a reinforcer assessment was conducted, during which each child was offered a chance to earn play time with each toy by completing academic work at one of five desks, with each associated with a different toy. Results indicated that the verbal forced-choice and observation methods predicted reinforcers with 70% and 60% accuracy, which was much higher than the nomination method (40%).

A follow up study was conducted with four elementary-age children with ADHD attending a summer school program (Northup, George, Jones, Broussard, & Vollmer, 1996). In this study, reward categories included edible items, tangible items, activities, escape from work, and teacher attention, and three preference assessment methods were evaluated. A reinforcer survey required the child to indicate how much they liked each reward category by selecting a rating according to a 3-pt, Likert-type t scale. A verbal forced-choice method, identical to Northup et al. (1995), was also presented. A pictorial forced-choice was the third method, and resembled the verbal method except that the child was required to physically pick between coupons representing the reward categories. The pictorial method resembled the procedures for forced-choice assessments that most often used with nonverbal participants. Next, a reinforcer assessment was conducted, during which the children were provided an opportunity to earn each reward
contingent on the completion of coding tasks. Following five sessions, visual inspection of the cumulative number of problems completed for each reward category determined those stimuli that actually functioned as reinforcers. Results indicated that the verbal and pictorial forced-choice methods accurately classified 70% and 80% of the reward categories, respectively, while survey accuracy was less accurate (55%). A post-administration of all three preference assessments, conducted 10 days later, indicated that the accuracy of the verbal forced-choice increased to 80%, but the two other methods remained the same.

Northup (2000) further evaluated the accuracy of the survey method for 20 students attending a summer camp for children with ADHD. The procedures for the survey and reinforcer assessment were similar to Northup et al. (1996), except that a fifth reward category (peer attention) was added, and the criterion for earning rewards was completion of easy math facts. Results indicated that the reinforcer survey was only 57% accurate in classifying reinforcers, suggesting that earlier findings were stable across a larger population, and with educationally relevant tasks.

Conclusions

The validity of preference assessments is well established for children and adults whose cognitive and language barriers prevent the use of verbal nomination or choice methods (Matson, Bielecki, Mayville, Smalls, Bamburg, & Bagliom, 1999; Didden & de Moor 2004). Relatively few studies have investigated preference assessments for children with age-appropriate verbal skills. This is unfortunate, for two reasons. First, using positive contingencies for work completion and accuracy is one of the most common approaches to dealing with various aspects of classroom management, including work accuracy (Stevens, Blackhurst, & Slaton, 1991), academic engagement (Narayan, Heward, Gardner, Courson, & Omness, 1990), and controlling problem behaviors (Charlop, Burgio, Iwata, & Ivancic, 1988). However, these programs may be less effective if rewards are arbitrarily selected. Second, the existing research in this area suggests that simply asking a child to nominate their favorite or preferred consequences may be ineffective, and is certainly less effective than more rigorous methods.

A review of the literature, however, reveals two methodological issues that may limit the generality of these findings to common practice. First, all of the previous studies
in this area have targeted tasks that were either non-academic (Northup et al., 1996) or non-challenging, “easy” academic work (Northup et al., 1995; Northup, 2000). Thus, it is unknown whether the relative differences between methods, or absolute levels of reinforcer accuracy, exist when children at-risk for academic challenges are confronted with difficult academic tasks. Second, in each of the three previous investigations the preference assessments were administered prior to the reinforcer assessment and, thus, there was no opportunity to evaluate whether the accuracy of the preference assessment methods improved over time and with repeated testing. This is an important limitation because behavior management plans are typically ongoing, and it is possible that a child – much like teachers, parents, and researchers - may learn more about their own preferences with sustained contact with contingencies. Northup et al. (1996) provided some evidence that the predictive utility of preference assessments are stable, but the follow-up phase was conducted 10 days after the final reinforcer assessment. The purpose of the current study was to replicate the procedures used by Northup et al. (1996) using more challenging academic work, and administering the preference assessments and reinforcer assessments repeatedly across time.
Method

Participants and Setting

Participants were 15 second grade children attending a four-week summer school program. The summer school was organized and staffed by faculty and graduate students in a school psychology program. All assessment activities pertaining to this study were conducted across consecutive days, during 30-minute one-to-one instructional settings. The remainder of the 3-hr school day included individual and small-group reading interventions, and large group independent seatwork. No performance contingencies were in place during other instructional times.

All children were selected to participate in the summer school by their school district’s Title I Coordinator. The primary factor in referral decisions was poor performance on the district’s spring administration of the Dynamic Indicators of Basic Literacy Skills (DIBELS; Good, Gruba, & Kaminski, 2002). Three children among the 15 (Nate, Kevin, Clay), however, were referred for behavioral concerns rather than reading fluency problems.

Screening Information

In order to evaluate their reading skills, children were administered three reading assessments during the first two days of the program.

Reading fluency. Fluency was assessed through curriculum-based measurement in reading (CBM-R; Shinn, 1989). Fluency was defined as the median number of words read correctly per minute from three 3rd grade passages. According to Good, Gruba, & Kaminski (2002), a beginning 3rd grade benchmark for fluency is 77 correct words per minute.

Reading comprehension. Comprehension was defined as the percentage of correct choices on a maze passage (Howell & Nolet, 2000). For this task, each child read a 3rd grade passage that was altered in the following manner: with the exception of the first and final sentence, every seventh word was replaced by a blank, and below each blank were three choices, only one of which was semantically and syntactically correct. To complete the maze, the child read the passage silently and circled the word that “made sense.” There was no time limit, and the child’s score was calculated as the percentage of
correctly circled choices. Howell and Nolet recommended that maze performance in the 60% - 80% accuracy range is considered within the child’s instructional level.

Oral reading and maze passages were randomly selected from a pool of 30 passages in each category. Each passage was a unique and complete story, approximately 200 – 250 words in length, and gauged to the appropriate readability level according to Microsoft Word Flesch-Kincaid readability formula. Children were administered 3rd grade materials to correspond with the grade level they would be entering in the subsequent school year.

General reading ability. To obtain a general measure of reading ability, the Woodcock Johnson III (WJ-III; Woodcock, McGrew, & Mather, 2001) Broad Reading scale was administered. The Broad Reading scale consists of three subtests: Letter-Word Identification, Reading Fluency, and Passage Comprehension. A scaled score for this domain (M = 100, SD= 15) was recorded for each child.

Dependent Measure

The dependent measure across all experimental conditions was coupon choice. During each assessment, coupons representing four categories of rewards were used. 
Award coupons represented “symbolic’ items such as certificates, stamps, or stickers that indicated a job well done. Prize coupons represented novelty items, including pencils, cards, and small toys. Note Home coupons represented brief written letters to parents, handwritten on school stationary, describing some positive behavior the child exhibited during the school day. Pizza coupons represented one piece of a pizza. With the exception of pizza, all coupons were exchanged for backup rewards on a one-to-one ratio at the end of each day, with separate plastic bins for reward type filled each morning. Pizza coupons were exchanged at the end of each week, with every five coupons earning a personal pan pizza delivered to the classroom. Each coupon had a unique color and image, was introduced on the first day of the program, and was utilized during preference and reinforce assessments. For the reinforcer assessment, a “control” coupon was added. This coupon was referred to simply as a “gray” coupon, contained no image, and could not be exchanged for any rewards.

Experimental Conditions
Nomination survey. In this condition, the child was read a nomination survey. Specifically, the examiner stated: “I am going to name some things that kids sometimes get in school. I want to know how much you like each of these things. After I name each thing, you tell me if you like it a little, a lot, or not at all.” Each reward category was named, and the child’s verbal response was recorded. A sample item was shown for prizes and awards, notes home were described, and for pizza, the child was told that “when you earn 5, you get a free personal pan pizza.” For this condition, any item rated “a lot” was classified as a preference. Thus, it was possible for no item, some items, or all four items, to be identified as a preference.

Verbal forced choice. During the verbal forced choice preference assessment, the examiner read 12 sentences that required the child to choose one of two reward categories (e.g., Would you rather earn a Prize...or an Award?). Each of the four reward categories was paired twice with each of the others, in reverse order of presentation. The examiner circled the child’s verbal response.

Using a verbal forced choice procedure, Northup et al. (1996) defined “high” preference as any item that was selected at least 75% of the times it was presented in a pair. In the current study, “moderate to high” preferences were targeted, and thus a criterion of 50% (i.e., an item was selected 3 of 6 times presented) was established. Using this criterion, it was possible for no item, some items, or all four items to be identified as a preference.

Reinforcer assessment. In this condition, the students participated in a reinforcer assessment. The reinforcer assessment was conducted as follows. First, the student was presented with one of five coupons (including the “control” coupon) and a maze passage, and was told:

“Here is a story with every seventh word missing. Circle the word that should go into the blank. If you make a mistake, you can go back and change your answer. For every seven correct choices, you will earn a ________ coupon. You can do as much as you want, as little as you want, or none at all. We will stop after 5 minutes, or if you say, ‘I’m done.’”

As the child worked, the examiner tallied the number of correct choices, and distributed a coupon each time the child correctly circled seven choices. After five minutes, or earlier, if the child said “I’m done,” the examiner placed a slash mark (/) after
the final completed choice, and presented the next coupon and the same directions. Once a maze passage was completed, a new one was begun. The reinforcer assessment continued until the child had an opportunity to work for each of the five coupons.

Due to its experimental nature, the reinforcer assessment was considered a “true” test of those stimuli that actually function as positive reinforcement. For this reason, the results of the reinforcer assessment were treated as the criterion variable in this study. Four to five complete trials were conducted for each participant over the course of the summer program, with coupons randomly ordered, and each trial was separated by a new administration of the nomination survey and verbal forced choice. For the reinforcer assessment, a reinforcer was defined summatively as any coupon for which the cumulative number of maze items completed exceeded the child’s mean performance across all coupons, including the control. This objective criterion was used primarily to avoid risky subjective interpretations of five alternating series across four to five measurement intervals. Equally important, this criterion allowed for the identification of no items (i.e., if all five, including the control, were equally effective), some items, or all items (i.e., if all four were equally effective versus the control).

The number of maze items completed was used to determine reinforcers, although the child actually received a coupon for every seven correct responses. This contingency was in place to ensure that the child did not arbitrarily circle answers in order to earn more coupons. The number of completed items, rather than correct items completed, was viewed as the most accurate estimate of the child’s motivation.

Reliability

The reliability of reinforcer assessment scoring was assessed for 55% of the maze passages by an independent observer, who scored each completed maze item as correct or incorrect. For each passage, the total number of agreements were divided by the number of agreements plus disagreements, and multiplied by 100%. The mean interobserver agreement (IOA) across all children and all maze passages was 95% (range, 72% to 100%).

Design and Procedures

A multi-element design was used to evaluate the reinforcement value of the five coupons during the reinforcer assessment. Each trial consisted of five experimental
conditions, or one opportunity to earn each coupon, and at least four trials were conducted for each child over the course of 12 instructional days. The nomination survey and verbal forced choice assessment was administered prior to each reinforcer assessment.
Results

Table 1 displays the screening information for each participant, and the results of their reinforcer assessment. The broad reading skills of most children, according to the WJ-III, were within the average range (M = 93), with two exceptions (Brianna and Megan). In terms of oral reading fluency, all but three children were below the literature-based benchmark of 77 (M = 55 CWPM). Administration of a maze passage during screening indicated that 9 of the 15 children performed within the instructional range of 60% to 80% recommended by Howell and Nolet (2000). Together, these data indicate that most of the participating children exhibited low average reading ability and substantial deficits in reading fluency, and that the 3rd grade maze passages were a sufficiently challenging reading task.

The results of the reinforcer assessment indicated that most children exhibited one of three patterns of responding. An example of each pattern is displayed in a cumulative graph in Figure 1. Four children, including Clay, earned each category during most trials (top panel). Nine children, including Patty, earned two to three categories during most trials (middle panel). For the final two children, including Nate, there was no clear differentiation between the five coupons, including the control (bottom panel). Although prize coupons were the favorite, no children consistently worked for a single category of coupons, at the exclusion of all others. Prize coupons were earned by 100% of the children, followed by pizza (75%), note home (33%), and award (27%).

Classification Accuracy

All children were administered at least four complete trials of the nomination survey, verbal forced choice, and reinforcer assessment. Classification accuracy for the two preference assessment procedures, at each trial, was evaluated in the following manner. According to criteria specified earlier, each reward category was classified as a preference or not (1 or 0, respectively) for the nomination and the verbal forced choice, and a reinforcer or not (1 or 0, respectively) for the reinforcer assessment. Next, total and corrected classification accuracy was calculated between each preference assessment method and the reinforcer assessment, using the entire population of rewards across the 4 categories and 15 children (N=60). Total accuracy was determined by calculating the sum of true positives and true negatives divided by the total possibilities. Table 2 displays the
total accuracy scores for the nomination and verbal forced choice. These data indicate that the classification accuracy of both methods were similar at each trial, and increased incrementally over time as the children became more familiar with the contingencies and backup rewards.

Given the high selection rates of pizza and prize, as both preferences and reinforcers, total accuracy may be inflated by chance agreement. Thus, Kappa (Cohen, 1960), which is a coefficient of nominal agreement beyond chance, was also calculated for each comparison. Table 2 also displays Kappa coefficients for the two preference assessments. These data indicate that the strength of association between the nomination survey and the reinforcer assessment was no different than chance during the first two trials, but emerged as significant during the third and fourth trial. On the other hand, Kappa coefficients for the verbal forced choice method were considerably higher at all four intervals. At all four intervals, the verbal forced choice method was consistently more accurate beyond chance, although total accuracy and the degree of improvement over time were relatively equal for both methods.

As noted by Northup (2000), all classification errors are not equal. False positives (i.e., accepting a reward that is not reinforcing) have far greater negative consequences for the child than false negatives (i.e., rejecting a reward that is reinforcing), as the success of a behavior plan depends more on the effectiveness of a reward that is used versus the ineffectiveness of a reward that is not used. Further analysis of each method, and a combination, was conducted to determine which errors are more likely, and which types of errors are reduced with repeated trials. Table 3 displays the frequency of classification errors on the fourth and final trial for the nomination survey, the verbal forced choice, and a combined approach. For the combined approach, a reward was classified as a preference only if it was identified as such by both methods. This table also displays the absolute improvement (e.g., increases in true positives, decreases in false positives) for these methods from the first to final trial.

Classification rates indicated that false positives were more likely for the nomination survey versus the verbal forced choice, which is a primary limitation of the survey format noted by Northup (2000). The improvement data, however, indicate that increased accuracy over repeated trials and exposure to contingencies was due primarily
to reductions in false positives (and subsequent increases in true negatives) for the nomination survey, while changes for the verbal forced choice were more equally distributed across all types of classification rates. A combined criterion decreased the frequency of false positives, but the overall accuracy (77%) was only marginally higher than the verbal forced choice alone. Functionally, these data indicate that the improved classification accuracy over time and exposure was due to the accurate re-classification of six to eight cases out of 60. The verbal forced choice improved by correctly classifying more reinforcers while also correctly rejecting more ineffective items. The nomination survey improved primarily by rejecting more ineffective items.
Discussion

Surveys are much maligned, but past research has neglected the most appealing element of the nomination survey method: They are relatively quick and efficient, and thus can be administered often, and with ease. Greater efficiency is not a virtue, however, unless they are reasonably accurate. This study examined the predicative validity of a nomination survey and forced choice assessment and sought to investigate whether the accuracy improved after repeated exposure to assessment and access to stimuli. One assumption of this study was that reinforcer identification in actual practice would typically be ongoing, particularly if educators were familiar with idiosyncratic effects of establishing operations and response effort.

Results indicated that verbal forced choice was superior to the nomination survey, particularly when chance agreements are ruled out. This supports earlier research (Northup et al., 1995; Northup et al., 1996; Northup 2000). Differences in accuracy were quite stable over time, with improved accuracy for both methods with each additional administration. By the third administration, both methods identified reinforcers beyond chance levels.

Implications

Results from this study maintain that verbal forced choice is the standard in stimulus preference assessments. This is the fifth demonstration of its superiority over nomination survey methods for verbal populations (Northup et al., 1995; Northup et al., 1996; Northup, 2000; Wilder et al., 2003). Forced choice remains the most predictive, effective, and accurate option for assessing reinforcers. This is true at the first or initial administration, and remains true for at least four consecutive trials. When compared to nomination survey, both methods become more predictive as familiarization with rewards and contingencies increased. However, differences still remained, with verbal forced choice continuing to be the best choice. At this point, nomination survey has not been shown to have any incremental value over verbal forced choice, nor has a combined approach utilizing both methods.

When examining only total accuracy, nomination survey and verbal forced choice methods had equivalent overall classifications, but when differences beyond chance were taken into consideration, the discrepancy between the two methods was quite large.
Unless chance agreement is considered, practitioners using nomination survey are likely to view it as equally effective. Kappa becomes a useful statistics in more stringent reporting of the effectiveness of a reinforcer assessment method because it evaluates differences beyond chance agreement. In order to elucidate these differences, Kappa measures should be reported in future research comparing stimulus preference assessment methods.

Limitations

This study utilized three categories of “tangible” stimuli and one “edible” category (i.e., pizza), using coupons that were cashed in after school. We selected these four reward categories because they were mutually exclusive, familiar to most students, and could be delivered in a manner that was least disruptive to the instructional schedule. Previous research has included a broader range of categories, including escape, teacher attention, and peer attention (Northup et al., 1996; Northup, 2000), which makes direct comparisons between studies more difficult. Different sampling of potential categories may be one reason why absolute accuracy levels were lower than previously reported. Thus, one limitation is that these findings cannot be generalized to other, potentially more robust categories of stimuli.

It should be noted, however, that the classification strength of these methods is most accurately reflected in the Kappa coefficient. Overall classification accuracy across trials and methods varied only between 62% and 75%. At first glance, these absolute scores might be dismissed as having relatively little practical value. When considering chance agreements, however, there is little question that verbal forced choice was superior to nomination survey, and that the relative improvement across trials was similar for both methods.

A second limitation of the current study is that only four trials were administered to all children. Thus, it is not clear whether the nomination survey method would eventually “catch up” to the forced choice method or if, indeed, both methods reached their maximum predictive strength after four trials.

A third limitation of the study is that the procedures did not identify or control for interaction effects. For example, a child might select a prize only if he or she had earned a criterion number of pizza coupons (i.e., enough for one pizza). Alternatively, a child
might work diligently for one note home, but not two, because two notes home may be perceived by the child as redundant. There were attempts to minimize satiation by offering several choices within each tangible category. For example, the award box had several different types of certificates, stickers, and other items of recognition. Anecdotally, there were no indications that children were aware of interaction effects.

A fourth limitation is that the study did not qualify low versus high preferences. This was intentional, as there are no universally accepted criteria for designating low versus high preference, which would render the qualification arbitrary. It is assumed that over time low preferences would improve (in terms of classification) at the same rate as high preferences, but there was no attempt to test this assumption in the current study.

A final limitation is the definition of reinforcer that was used in the study. For purposes of analysis, a true reinforcer was defined as a reinforcer for which the student completed a number of maze items that exceeded the child’s mean performance across all coupons, including the control. A potential alternative would be basing the definition of a reinforcer on visual inspection of graphic representations of the child’s completion of maze items for each category over time. Although visual inspection is reliable for detecting robust differences, and could possibly been used to identify the “most” effective reinforcer, we developed a summary statistic that would allow equal degrees of freedom across all three assessments. In other words, it was important that we used a standard that would allow for all coupons (or none at all) to be classified as a reinforcer. In nearly every case, visual inspection proved quite useful for distinguishing the relative effects of each coupon over time. It was much more difficult, however, to use visual inspection to make absolute judgments as to whether each coupon was a reinforcer.

**Future Directions**

One important issue is that the current investigation, like all previous studies of stimulus preferences, did not directly evaluate the impact of establishing operations. An establishing operation is an event, condition, or state that makes a reinforcer more or less potent at a given time (Miltenberger, 2004). For example, if a child has not eaten all day, an edible reinforcer will be more potent than if the child has just finished a large, satisfying meal. Hunger, or the amount of time between meals, is the establishing operation in this scenario. Other known establishing operations include immediacy and
novelty. Knowledge on reinforcement assessment methods can only go so far until establishing operations are addressed. Future research addressing establishing operations would require a highly controlled research environment, which could be difficult to create in an academic setting.

Another direction for future research would be to add a category of negative reinforcement. Many students are reinforced by items such as escape. Examples of this could include free time from instruction and homework passes to get out of assignments. These types of reinforcers are commonly used in classrooms and thus their inclusion in a future study could yield results more likely to generalize to the classroom environment.

A third direction for future research lies in matching students to their ability levels. All of the children in the current study completed comprehensive maze passages at the third grade level. According to baseline data, some of the children were working at their instructional level and some were working at their frustrational level. Future research could manipulate difficulty level to determine if stimulus preference assessment methods are more accurate at one level versus another.

Finally, there is a need to investigate the reinforcement value of “choice” itself. In a typical reinforcer assessment, children are given a choice among various rewards, and the primary focus is on determining which rewards actually function as reinforcers. There are rare instances, in “real” life, when a target response actually has several competing consequences. More often than not, a child’s academic productivity or social actions have a single consequence, such as teacher approval or peer avoidance. Thus, it is reasonable to assume that simply providing choice may have some inherent value to children, and that this value may be even more powerful than any one particular reward or reward category. Future work in this area should further investigate, for example, whether a high preference is a more effective reinforcer than choice among low preference items.

**Conclusions**

It has been established that one of the most popular and effective methods of classroom management is the use of positive contingencies (Charlop et al., 1998; Narayan et al., 1990; Stevens et al., 1991). With so many reward systems in use, it is essential that effective reinforcers be identified. Reinforcing a student with an item that will not change behavior can only result in frustration for the teacher and the student.
Fortunately, the forced choice assessment has been shown multiple times to be effective at identifying reinforcers (Northup et al., 1995; Northup et al., 1996; Northup, 2000). The current study also established that the forced choice assessment becomes a better predictor of reinforcers when used repeatedly with the same student. The forced choice assessment only involves slightly more effort to use than the nomination survey and delivers significantly more powerful results. When effective reinforcers are identified and used to improve student performance, everyone wins.
References


Weaver, A. D. (2004). *Idiosyncratic effects of different classroom rewards on task performance and on-task behavior*. Thesis (Ph.D.)--Mississippi State University. School of Psychology in the Department of Counselor Education and Educational Psychology.


Table 1. Descriptive data and reinforcer assessment results for each participant.

<table>
<thead>
<tr>
<th>Child</th>
<th>Broad Reading</th>
<th>CBM</th>
<th>Maze</th>
<th>Reinforcer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brianna</td>
<td>73</td>
<td>19</td>
<td>38%</td>
<td>Prize Pizza Note Award</td>
</tr>
<tr>
<td>Clay</td>
<td>105</td>
<td>111</td>
<td>77%</td>
<td>Prize Pizza Note Award</td>
</tr>
<tr>
<td>Chelsey</td>
<td>105</td>
<td>71</td>
<td>73%</td>
<td>Prize Pizza Note</td>
</tr>
<tr>
<td>Erin</td>
<td>89</td>
<td>46</td>
<td>96%</td>
<td>Prize Pizza Note</td>
</tr>
<tr>
<td>Nate</td>
<td>100</td>
<td>86</td>
<td>80%</td>
<td>Prize Award Control</td>
</tr>
<tr>
<td>Kevin</td>
<td>105</td>
<td>89</td>
<td>88%</td>
<td>Prize Note Award</td>
</tr>
<tr>
<td>Mike</td>
<td>89</td>
<td>65</td>
<td>62%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Sarah</td>
<td>94</td>
<td>29</td>
<td>69%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Caren</td>
<td>92</td>
<td>36</td>
<td>69%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Megan</td>
<td>72</td>
<td>26</td>
<td>45%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Kristi</td>
<td>98</td>
<td>50</td>
<td>68%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Stephanie</td>
<td>92</td>
<td>51</td>
<td>68%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Ashley</td>
<td>95</td>
<td>67</td>
<td>59%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Patty</td>
<td>92</td>
<td>27</td>
<td>69%</td>
<td>Prize Pizza</td>
</tr>
<tr>
<td>Laura</td>
<td>96</td>
<td>49</td>
<td>54%</td>
<td>Prize Award</td>
</tr>
</tbody>
</table>

Table 2. Accuracy and kappa coefficients for nomination and verbal forced choice.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Survey Accuracy</th>
<th>Survey Kappa</th>
<th>Verbal Forced Choice Accuracy</th>
<th>Verbal Forced Choice Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62%</td>
<td>.09</td>
<td>63%</td>
<td>.22</td>
</tr>
<tr>
<td>2</td>
<td>67%</td>
<td>.22</td>
<td>70%</td>
<td>.37**</td>
</tr>
<tr>
<td>3</td>
<td>70%</td>
<td>.32**</td>
<td>72%</td>
<td>.42**</td>
</tr>
<tr>
<td>4</td>
<td>72%</td>
<td>.35**</td>
<td>75%</td>
<td>.48**</td>
</tr>
</tbody>
</table>

Note: * = p < .01; ** = p < .001
Table 3. Frequency of classification rates and improvement across four trials for the two preference assessment procedures ($N=60$).

<table>
<thead>
<tr>
<th></th>
<th>True Positives</th>
<th>True Negatives</th>
<th>False Positives</th>
<th>False Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nomination Survey</td>
<td>33</td>
<td>10</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>+1</td>
<td>+5</td>
<td>-5</td>
<td>-1</td>
</tr>
<tr>
<td>Verbal Forced Choice</td>
<td>29</td>
<td>16</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>+3</td>
<td>+4</td>
<td>-4</td>
<td>-3</td>
</tr>
<tr>
<td>Combination</td>
<td>27</td>
<td>19</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>+2</td>
<td>+6</td>
<td>-6</td>
<td>-2</td>
</tr>
</tbody>
</table>
Figure Caption

*Figure 1*: An example of three types of response patterns observed during the reinforcer assessment.