

Running head: Quantifying Error Patterns

Quantifying Errors of Bias and Discriminability in Conditional-Discrimination Performance in Children Diagnosed with Autism Spectrum Disorder

Courtney Hannula<sup>1</sup>, Corina Jimenez-Gomez<sup>2</sup>, Weizhi Wu<sup>1</sup>, Adam T. Brewer<sup>3</sup>, Tiffany Kodak<sup>4</sup>, Shawn P. Gilroy<sup>5</sup>, Blake A. Hutsell<sup>6</sup>, Brent Alsop<sup>7</sup>, and Christopher A. Podlesnik<sup>2</sup>

<sup>1</sup> Florida Institute of Technology and The Scott Center for Autism Treatment

<sup>2</sup> Auburn University

<sup>3</sup> Western Connecticut State University

<sup>4</sup> Marquette University

<sup>5</sup> Louisiana State University

<sup>6</sup> East Carolina University

<sup>7</sup> Otago University

Author Note

The first experiment was conducted in partial fulfillment of the first author's Masters thesis. Thanks to Dr. Darby Proctor for helpful suggestions when designing these experiments and to Tiara Putri and Carolyn Ritchey for their assistance in conducting this research. Please address correspondence to Chris Podlesnik ([cpodlesnik@auburn.edu](mailto:cpodlesnik@auburn.edu)).

### Abstract

Antecedent- and consequence-based procedures decrease errors during conditional discrimination training but are not typically guided by error patterns. A framework based in behavioral-choice and signal-detection theory can quantify error patterns due to (1) *biases* for certain stimuli or locations and (2) *discriminability* of stimuli within the conditional discrimination. We manipulated levels of disparity between sample (Experiment 1) and comparison (Experiment 2) stimuli by manipulating red saturation using an ABA design with children diagnosed with autism spectrum disorder (ASD). Lower disparities decreased discriminability and biases were observed for some participants during the low-disparity conditions. These findings demonstrate the use of these analyses to identify error patterns during conditional-discrimination performance in a clinically relevant population under laboratory conditions. Further development of this framework could result in the development of technologies for categorizing errors during clinically relevant conditional-discrimination performance with the goal of individualizing interventions to match learner-specific error patterns.

*Keywords:* autism spectrum disorder, bias, children, conditional discrimination, discriminability, errors, matching to sample

The ability to discriminate among stimuli is foundational for many skills, including academics, socializing, communicating, engaging in self-care routines, and others (Green, 2001). Simple discriminations involve three terms (Davison & Nevin, 1999), an antecedent, a behavior, and a consequence. For example, an adult tells a child to sit down and provides praise once the child is seated. In contrast, conditional discrimination involves a four-term contingency in which a conditional stimulus (e.g., the printed word “dog”) changes the function of a stimulus in the comparison array (e.g., the picture of the dog becomes the  $S^D$ , or  $S+$ , and the picture of the cat becomes the  $s$ -delta, or  $S-$ ). Pointing to a picture of a dog produces reinforcement in the presence of the printed word “dog” and extinction in the presence of the printed word “cat” and vice versa.

Conditional discriminations typically are taught using matching-to-sample (MTS) procedures in which the conditional stimuli, or sample stimuli, are comprised of two or more visual or auditory stimuli (e.g., the printed words “dog” and “cat”). Correct responses involve choosing the comparison stimulus (e.g., the picture of the dog or cat) that corresponds with the sample stimulus, which typically results in the delivery of a reinforcer. Errors involve selecting the nonmatching or noncorresponding comparison that would typically produce extinction or punishment (e.g., Fisher, Pawich, Dickes, Paden, & Toussaint, 2014). Conditional discriminations have been taught in laboratory studies with both humans and nonhumans (see Beran, Menzel, Parrish, Perdue, et al., 2016; Davison & Nevin, 1999, for reviews) and to individuals diagnosed with autism spectrum disorder (ASD) and other developmental disabilities (e.g., Fisher, Retzlaff, Akers, DeSouza, Kaminski, & Machado, 2019; McIlvane, Kledaras, Gerard, Wild, & Smelson, 2018; Williams, Johnston, & Saunders, 2006).

Although there is extensive research using MTS procedures to train conditional discriminations, some individuals in both clinical and laboratory settings exhibit persistent errors

during training (see Green, 2001; Dube & McIlvane, 1997; Grow, Carr, Kodak, Jostad, & Kisamore, 2011; Saunders & Spradlin, 1989). Both antecedent and consequent manipulations have been developed to decrease errors. Antecedent manipulations arrange stimuli during the beginning of trials to increase contact with the relevant stimuli. For one example, differential observing responses require participants to emit different responses depending on the currently presented sample stimulus (e.g., Dube & McIlvane, 1999; Fisher, Kodak, & Moore, 2007; Fisher et al., 2019; Truppa, Mortari, Garofoli, Privitera, & Visalberghi, 2011). Fisher et al. (2019) had participants echo the auditory sample stimulus prior to selecting a picture in the comparison array. In contrast, consequent manipulations introduce interventions following errors. Examples of such procedures include repeating trials following errors (Da Silva Barros, De Faria Galvao, & McIlvane, 2002; McGhan & Lerman, 2013), punishment (Fisher et al., 2014), and experimenter modeling of the correct response (e.g., Kodak et al., 2016).

Although previous studies have used antecedent- and consequence-based interventions, there are two related problems with their implementation. First, few studies describe a method for selecting interventions in response to specific error patterns in performance. Although research comparing the efficacy and efficiency of antecedent- and consequence-based interventions identified specific procedures as most efficacious and efficient for particular learners (e.g., Cubicciotti, Vladescu, Reeve, Carroll, & Schnell, 2019; Grow et al., 2011; Kodak, et al, 2016; Kodak, Fisher, Clements, Paden & Dickes, 2011; McGhan & Lerman, 2013), these strategies are infrequently implemented in response to specific error patterns during training. Second, the identification of interventions based on learner errors requires strategies to identify and distinguish between the types of errors that may occur during conditional discrimination

training. However, there is a paucity of research on strategies to identify and characterize learner errors.

In a notable exception, Grow et al. (2011) categorized errors during auditory-visual conditional discrimination training as molar win-stay errors, molecular win-stay errors, and errors from position biases. Win-stay errors (Lovaas, 2003; Kangas & Branch, 2008; Williams et al., 2006) reflect the continued selection of comparisons reinforced during previous trials (molecular) or previous training conditions (molar). Position biases reflect the continued selection of a location in the comparison array regardless of the sample. Errors based on discriminability of experimental stimuli were not categorized directly but inferred when win-stay and position errors did not account for imperfect performance. Although the methods of Grow et al. offer a beneficial starting point for categorizing errors during conditional-discrimination training, strategies to characterize and mitigate errors on an individual basis remain largely unexplored, and these could provide a useful tool for increasing the efficiency of conditional-discrimination training.

Davison and Tustin (1978) developed a quantitative framework based in behavioral-choice and signal-detection research to identify two error patterns exhibited during conditional discriminations – errors of *discriminability* versus errors of *bias*. Discriminability is a measure resulting from the degree to which relevant sample or comparison stimuli are distinguishable features of conditional discrimination (see Davison & Nevin, 1999, for a relevant discussion). For an example manipulating sample-stimulus disparity from basic laboratory research, Davison and McCarthy (1987) trained pigeons to peck the right key to access reinforcer delivery if the center key was illuminated for 5 s and to peck the left key to access reinforcer delivery if the center key was illuminated for 12 other durations. Accuracy was lower when the disparity

between sample durations on the center key was small (e.g. 5 s vs. 7.5 s) than when greater (e.g., 5 s vs 57.5 s). For example, samples with lower disparity such as “dog” and “log” would likely produce poorer accuracy than “dog” and “stick.” Relatedly, Stromer, McIlvane, Dube, and Mackay (1993) demonstrated that presenting multiple samples during trials resulted in more difficult sample discriminations as indicated by reductions in accuracy during MTS procedures in children diagnosed with developmental disabilities (see also Zentall, 2005, for a review of related research with laboratory animals). Thus, decreases in accuracy with more similar sample stimuli (i.e., lower disparity) are predicted to be reflected in decreases in discriminability.

In contrast, accuracy can also decrease under situations in which stimuli are not necessarily impossible to discriminate but biases compete with responding accurately. Specifically, bias is a measure resulting from the degree of preference for a comparison stimulus or location. For example, Cumming and Berryman (1961) documented inherent biases for both stimulus color and stimulus location in pigeons because no obvious variables accounted for the biases. Pecking the comparison key matching the color from the sample produced delivery of food reinforcers. All pigeons exhibited more frequent responding on one comparison location regardless of where the correct comparison stimulus appeared – a location bias. Furthermore, a few pigeons also demonstrated a stimulus bias by choosing a particular color comparison more frequently. Similarly, in applied research with typically developing children, Schneider, Devine, Aguilar, and Petursdottir (2018) reported both stimulus and location biases in MTS tasks presenting birds, flowers, or flags. Further, differences in reinforcement variables for accurate performance also can impact bias. For example, Alsop et al. (2016) arranged higher versus lower probabilities of reinforcement for accurate matches between sample-comparison pairs in children diagnosed with attention-deficit hyperactive disorder. They observed consistent biases toward

comparison stimuli resulting in more likely reinforcement for accurate matches. Thus, differences in reinforcer variables are predicted to be reflected in biases but inherent preferences can also influence biases.

Davison and Tustin's (1978) quantitative framework offers potential benefits for applied researchers and clinicians because it can identify patterns of errors due to discriminability alone or biases for comparison stimuli or locations. Quantifying these error patterns could contribute to identifying the source(s) of errors contributing to poor conditional-discrimination performance. Refer to Figure 1 as we describe the quantification of discriminability and bias through the distribution of correct responses (11, 22) and errors (12, 21) at the comparison choice points of conditional discriminations. For example, the written word dog as the sample ( $S_1$ ) followed by a choice of the comparison picture of the cat ( $C_2$ ) would be  $Error_{12}$ .

According to Davison and Tustin (1978), discriminability can be quantified with  $\log d$ :

$$\log d = .5 \log \left[ \left( \frac{Correct_{11}}{Error_{12}} \right) \left( \frac{Correct_{22}}{Error_{21}} \right) \right], \quad (1)$$

where *Correct* and *Error* refer to correct (11, 22) and error (12, 21) responses, respectively (see Hutsell & Banks, 2015, 2017; Shahan & Podlesnik, 2006, 2007). Note that all logarithms throughout are base 10.  $\log d$  measures the accuracy of the choice pattern (theoretically) independently from the reinforcer distribution or any inherent biases the individual exhibits (Alsop & Rowley, 1996; Davison & Jenkins, 1985). Values of  $\log d$  range from negative to positive infinity, although  $\log d$  is zero at chance performance with equal correct and error responses to comparisons, and increases as accuracy improves. Thus,  $\log d$  quantifies discriminability between sample stimuli or between comparison stimuli, with larger positive values indicating greater discriminability (Davison & McCarthy, 1987). Given the  $\log_{10}$  space,

obtained  $\log d$  values of 1 indicate a 10:1 ratio of accurate-to-inaccurate responses, obtained  $\log d$  values of 2 indicate a 100:1 ratio, etc.

According to Davison and Tustin (1978), biases between comparison stimuli and location are quantified by the equations for  $\log b$  and are (theoretically) independent from  $\log d$  (discriminability). Equation 2 quantifies bias between comparison stimuli using terms as they appear in Equation 1:

$$\log b (\text{stimulus}) = .5 \log \left[ \left( \frac{\text{Correct}_{11}}{\text{Error}_{12}} \right) \left( \frac{\text{Error}_{21}}{\text{Correct}_{22}} \right) \right]. \quad (2)$$

Equation 3 quantifies bias for comparison location and is calculated similarly:

$$\log b (\text{location}) = .5 \log \left[ \left( \frac{\text{Correct}_{\text{left}}}{\text{Correct}_{\text{right}}} \right) \left( \frac{\text{Error}_{\text{left}}}{\text{Error}_{\text{right}}} \right) \right]. \quad (3)$$

In Equation 3,  $\text{Correct}_{\text{left}}$  and  $\text{Correct}_{\text{right}}$  refer to the correct responses to the comparison stimuli and  $\text{Error}_{\text{left}}$  and  $\text{Error}_{\text{right}}$  refer to incorrect responses to comparison stimuli given the sample-stimulus presentation (see Jones & White, 1992). Therefore, these two  $\log b$  equations measure the degree to which the individual emits more responses to one comparison stimulus relative to the other comparison stimulus (Equation 2) or one comparison location relative to the other comparison location (Equation 3).  $\log b$  for both stimulus and location range from negative to positive infinity, with  $\log b$  values of zero denoting no indication of bias to a particular comparison stimulus or location. Given the  $\log_{10}$  space, obtained  $\log b$  values of  $\pm 1$  indicate a 10:1 or 1:10 ratio, respectively, of accurate-to-inaccurate responses; obtained  $\log d$  values of  $\pm 2$  indicate a 100:1 or 1:100 ratio, etc.

Despite the potential clinical usefulness of quantifying error patterns using Davison and Tustin's (1978) framework (McCarthy, 1991),  $\log d$  and  $\log b$  have mostly been used to categorize errors in conditional-discrimination performance in laboratory research (see Davison and Nevin, 1999, for a review). The only application of this general framework comes from



Fisher et al. (2014) in children diagnosed with developmental disabilities, who hypothesized that introducing punishment following errors would increase the discriminability of the contingencies in effect during the choice between comparisons. Though conceptually consistent with the general approach, Fisher et al. did not use the quantitative framework in their analyses. Due to this general lack of quantitative analysis of error patterns during conditional discriminations in applied research, the overall goal of the present research is to initiate the development and use of this quantitative framework for characterizing errors in conditional-discrimination performance with clinically relevant populations.

The present study arranged two experiments assessing conditional-discrimination performance in children diagnosed with ASD on an automated touchscreen interface. Experiment 1 manipulated the disparity (i.e., similarity) of sample stimuli, and Experiment 2 manipulated the disparity of comparison stimuli across successive conditions. A number of obstacles exist to applying these analyses using clinically relevant stimuli at this stage, providing justification for the laboratory approach. Clinically relevant stimuli necessarily are individualized, resulting in stimuli differing qualitatively within and among individuals, as well as in complexity, disparity, prior exposure, and salience. Validating these analyses with well-controlled stimuli will provide the necessary foundation upon which to investigate methods for manipulating the disparity of clinically relevant stimuli thereafter. Therefore, both experiments manipulated stimulus disparity by changing color saturation (i.e., lighter to darker reds) from a large disparity to a small disparity, followed by a return to the large disparity according to an ABA design. We calculated percentage correct,  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location) throughout all conditions to characterize how changing similarity of the sample and comparison stimuli impacted accuracy and specific error patterns. Furthermore, the manipulation of disparities simulate changes in task

difficulty that are a natural part of teaching conditional discriminations clinically and in school settings.

## Experiment 1

### Method

**Participants.** Alfred, Harry, and Suzie participated in this study. All participants were recruited from a center offering early intensive behavioral intervention (EIBI) services to young children diagnosed with ASD. Participants all demonstrated the ability to follow simple instructions, sit or stand for five-min sessions, and emit the gross-motor response of pressing the touchscreen device. During consent meetings, all parents reported no diagnoses of color blindness.

Alfred was six years old and had been receiving EIBI services intermittently for three years with continuous service for the last 15 months. He was diagnosed with ASD, Unspecified Disruptive Impulse-Control and Conduct Disorder, Stereotypic Movement Disorder with Self-Injury, and Phonological Disorder. His score on the *Verbal Behavior Milestones Assessment and Placement Program* (VB-MAPP; Sundberg, 2008) was consistent with Level 3 (i.e., 30-48 months old) and he obtained 15 out of 15 on the Visual Performance/MTS section. Harry was four years old, diagnosed with ASD, and had been receiving EIBI services for eight months. His score on the VB-MAPP was consistent with Level 3, with a score of 11 out of 15 on the Visual Performance/MTS section. Suzie was four years old at the beginning of the study, diagnosed with ASD, and had been receiving EIBI services for 11 months. Her score was consistent with Level 3 on the VB-MAPP, with a score of 14.5 out of 15 on the Visual Performance/MTS section.

**Setting and Materials.** Sessions were conducted in a small room at the university EIBI center. Each room contained a table and chairs, edibles, a video camera, and a touchscreen Windows®-based laptop with a 21.7 cm by 13.6 cm screen with sessions programmed using Paradigm® software (Factari, 2018). On the touchscreen, colors of the background, sample stimuli, and comparison stimuli were defined by RGB color values supported in all browsers. Training stimuli were blue (R13G1B255) or yellow (R255G255B0) for samples and comparisons. Experimental stimuli were light pink (R255G155B155), pink (R255G100B100), dark pink (R255G95B95), light red (R255G70B70), red (R255G51B51), and dark red (R188G0B0). The background was black (R0G0B0) throughout all sessions to guard against negative afterimages. Sample and comparison stimuli were 4.4 cm by 4.1 cm. Samples were 8.6 cm from the left and right sides of the screen and 0.5 cm from the top. Comparisons were 4.8 cm from the top and bottom of the screen, 0.5 cm from the nearest side, and 12.9 cm from one another.

### **Response Measurement**

The primary dependent measures were the number of correct responses and errors during each session. Four response types were recorded for correct responses and errors to the red and pink comparison stimuli. Correct responses were defined as touching the comparison stimulus matching or corresponding with the sample stimulus. Errors were defined as touching the comparison stimulus that did not match or correspond with the sample stimulus. The location (e.g., left or right) of correct and incorrect responses was also collected. We also recorded the number of missed trials in which a participant did not respond within 30 s of the sample or comparison presentations, although missed responses were not included in the analyses.

Correct responses and errors were then analyzed using Equations 1, 2, and 3 (see Supporting Information for examples of these analyses). We analyzed individual-participant data from both individual sessions and, to summarize terminal performance within conditions, aggregated across the final seven sessions of each of the three experimental conditions. Aggregation of the final sessions across conditions used Equations 1, 2, and 3 with summed values of correct responses and errors across all seven sessions in the respective condition. We included seven sessions of data to equate data counts entered into Equations 1, 2, and 3 and because this was the fewest number of sessions needed to complete one phase across all participants (Shahan & Podlesnik, 2006, 2007).

In Equations 1 and 2,  $Correct_{11}$  and  $Correct_{22}$  refer to choosing the darker or lighter comparison following the darker or lighter sample, respectively.  $Error_{12}$  and  $Error_{21}$  refer to choosing the lighter or darker comparison following the darker or lighter sample, respectively. In Equation 3,  $Correct_{left}$  and  $Correct_{right}$  refer to choosing the left comparison or right comparison, respectively, when it corresponded with the sample stimulus. In contrast,  $Error_{left}$  and  $Error_{right}$  refer to choosing the left comparison or right comparison, respectively, when it did not correspond with the sample stimulus.

Percent correct was calculated by dividing the total number of correct responses by the total number of completed trials for the session and multiplying by 100. Equations 1, 2, and 3 cannot be calculated with a zero value in one of the four response categories. Therefore, we added a constant (0.25) to each response category when calculating  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location), as described previously (e.g., Alsop, 2004; Brown & White, 2005; Hautus, 1995). With 24 trials per session and the 0.25 added correction to each cell of Equations 1, 2, and 3, the minimum and maximum  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location) were  $\pm 1.69$  if the

numerators or the denominators were zero values. Arithmetically,  $\pm 1.69$  is a range of 47:1 to 1:47.

### **Preference Assessment**

Small pieces of preferred edibles served as primary reinforcers. Caregivers and clinicians nominated putative highly preferred edibles for each participant. Before each session, the experimenter conducted a brief multiple-stimulus-without-replacement (MSWO) preference assessment (Carr, Nicolson, & Higbee, 2000). The same choices were displayed in each MSWO for the remainder of the study. The first two edibles selected during the MSWO were randomly selected and delivered after each correct response along with praise and a star.

### **Procedures**

Depending on availability, participants attended experimental sessions two to three times per week, with two to five sessions being conducted per visit. All training and experimental sessions consisted of 24 trials, with the exception that training consisted of 25 trials per session for Harry only. The 24 trials per session comprised of 12 presentations of each sample with the order and location of the sample and comparison stimuli counterbalanced in a predetermined list that was shuffled prior to the beginning of each session. For Harry's training, the list restarted after 24 trials and he received one additional presentation of one trial type. Reinforcer deliveries were comprised of preferred edibles presented by the experimenter, along with praise (e.g., "You got it!") and a 3-s presentation of a 5 x 5 cm star image positioned 8.2 cm from the sides and 4.2 cm from the top and bottom, followed by a 2-s black screen. Errors were followed by a 5-s black screen only. The next trial initiated following the offset of the black screen. Across participants, session duration averaged 2.9 min (Range: 1.9-5.2 min).

**Training.** An experimenter taught participants to respond during the MTS procedure on the touchscreen device. Participants were first exposed to a blue or yellow sample stimulus. Touching the sample once removed the sample and presented one identical comparison stimulus on one side of the touchscreen. Touching the comparison stimulus resulted in a reinforcer delivery. At the beginning of training, the participant was instructed to “do this” with a model or physical prompt as needed. A most-to-least prompting strategy (e.g., full physical, partial physical, tap, gesture) was used to fade prompts (MacDuff, Krantz, McClannahan, 2001).

Once accurate responding occurred reliably and independently for 95% of trials or higher across two consecutive sessions, two comparison stimuli were presented following sample presentations – one comparison stimulus was identical to the sample. Choosing the correct, identical comparison resulted in reinforcer delivery (edible and praise) and touching the incorrect, non-identical comparison resulted in the black screen. Our original criterion for participants to begin experimental sessions was independent correct responses at or above 90% for two consecutive sessions. However, accuracy for Alfred and Harry did not reliably increase above 90% despite being reliably above chance. Therefore, we began experimental sessions with Alfred and Harry once responding occurred independently and we determined percent correct to be stable. Alfred, Harry, and Suzie were exposed to nine, thirteen, and four sessions, respectively, of training with the MTS procedure before beginning experimental sessions (data not shown but available from last author upon request).

**Experimental sessions.** The procedural arrangement was similar to training sessions with the exception of including two comparisons in the array, providing no prompts, changing the color of the stimuli, and the inclusion of non-identical MTS trials. The samples during Phases

1 and 3 had greater visual disparity than during Phase 2. The effects of sample disparity on responding were evaluated within an ABA experimental design.

**Phases 1 and 3: High disparity.** The sample stimuli were light pink and dark red, while the comparison stimuli were pink and red (see Supporting Information). Correctly touching the pink comparison after the light pink sample and touching the red comparison after the dark red sample resulted in reinforcer delivery. Errors only produced the black screen. Following Phase 1, Phase 2 began once accurate responding reached stability with no increasing or decreasing trends using visual inspection (Sidman, 1960). Following Phase 2, Phase 3 began once responding reached stability again.

**Phase 2: Low disparity.** Sample stimuli were more similar (lower disparity) compared with Phase 1. Sample stimuli were the dark pink and light red (see Supporting Information). Comparison stimuli were identical with those in Phase 1. Correctly touching the pink comparison after the dark pink sample and touching the red comparison after the light red sample resulted in reinforcer delivery. Errors only produced the black screen.

### **Procedural Integrity**

We assessed procedural integrity for delivery of edibles during presentations of the star onscreen during reinforcer deliveries for 33% of sessions. Procedural integrity was assessed for each session by dividing the total number of trials implemented with integrity by the total number of trials in a session, and then converting the result to a percentage. Procedural integrity was 100% for Alfred and Sally and 99% (range 96-100%) for Harry.

### **Analytical Plan**

As a complement to visual analysis, error patterns were also evaluated using a linear mixed-effects (i.e., multi-level) modeling approach (DeHart & Kaplan, 2019). Briefly, mixed-

effects models are often used in various fields (e.g., ecology) to answer questions when data fail to meet the assumptions (i.e., absence of outliers) of more commonly used tests (e.g., Analysis of Variance). This approach has been increasingly applied to the time-series data included in single-case research designs, even with small groups of participants – see DeHart and Kaplan (2019) for a demonstration with humans and Nall et al. (2019) for a demonstration with non-humans.

Although presented and interpreted here, the goal of these analyses were exploratory and to obtain measures of effect that would support larger, more expanded trials in the future.

Specifically, this modeling approach was to ascertain the effects of disparity on each of the  $\log d$ ,  $b$  (stimulus), and  $b$ (location) measures. In each of these separate analyses of the level of disparity (i.e., High, Low) was entered as a fixed effect. Although simpler methods could be applied, comparisons using mixed-effects avoid the undesirable compression of individual variability into singular means or ranks across phases or groups (DeHart & Kaplan, 2019). Study analyses were performed using the R Statistical Program (R Core Team, 2018) using the *lme4* package (Bates et al., 2015). In separate analyses of  $\log d$ ,  $b$  (stimulus), and  $b$  (location) the level of disparity (i.e., High, Low) was included as a fixed effect (Phase) and random effects (i.e., varying intercepts, slopes) were included as appropriate following the results of likelihood ratio tests.

## Results and Discussion

Figure 2 shows percent correct,  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location) for Alfred, Harry, and Suzie across successive sessions of the High, Low, and High Disparity conditions. In the top row of Figure 2, the percent correct was higher during the High Disparity conditions than during the Low Disparity conditions for all participants. Table 1 shows that reinforcers per session also were higher during the High Disparity conditions than during the Low Disparity conditions for all participants. The lower obtained reinforcers per session during the Low



Disparity condition is expected because the reinforcers are delivered according to FR 1 schedules for accurate matches.

Decreases in percent correct show a decrease in performance with lower sample disparity but do not indicate whether the increase in errors resulted from decreases in discriminability only or the development of stimulus or location biases. In the second row of Figure 2,  $\log d$  (i.e., discriminability) also was higher during the High Disparity conditions than during the Low Disparity conditions for all participants. Thus, changes in sample disparity across phases resulted in anticipated changes in  $\log d$  – low sample disparity reduced discriminability. Stated another way, discriminability was lower when the samples appeared more similar.

The third and bottom row of Figure 2 show bias as  $\log b$  (stimulus) and as  $\log b$  (location), respectively. Zero values along the y-axis indicate no bias. Positive values for  $\log b$  (stimulus) in the third row indicate more responses toward Comparison Stimulus 1 than 2 and negative values indicate more responses toward Comparison Stimulus 2 than 1. Positive values for  $\log b$  (location) in the bottom row indicate more responses toward the left comparison stimulus than right and negative values indicate more responses toward the right comparison stimulus than left. We included gray bands to signal the range of values within which errors are unlikely to reflect patterns of bias but, instead, are more likely to reflect general patterns of variability. The gray bands ranging  $\pm 0.368$  in log units (arithmetically, 2.3:1 to 1:2.3) indicate the range in which values would appear if produced by a single error across all 24 trials (e.g., Alfred's final six sessions). In such cases, overall accuracy can remain relatively high despite bias values being nonzero. Many other combinations of errors could also fall within the gray bands but would nevertheless reflect only minor deviations from zero bias. In addition, occasional values falling outside these bands likely reflect only general patterns of variability in

behavior rather than reliable patterns of bias. In contrast, values reliably falling outside these bands would indicate patterns of bias, especially when accompanied by low levels of accuracy.

In the third row of Figure 2,  $\log b$  (stimulus) typically fell within the gray bands and only occasional individual values extended beyond the gray bands (e.g., Sessions 17 and 19 for Suzie). Therefore, no reliable stimulus bias appeared in most conditions across participants. In the Low Disparity condition for Harry only, in contrast,  $\log b$  (stimulus) was negative and extended beyond the gray bands for 8 of 12 sessions. This pattern of negative  $\log b$  (stimulus) values indicate a bias for Comparison Stimulus 2 when the sample disparity was low. Thus, Harry's performance in the Low Disparity condition show that decreases in sample disparity certainly impact discriminability, as shown by  $\log d$  above, but decreases in disparity also can produce bias.

In the bottom row of Figure 2,  $\log b$  (location) typically fell within the gray bands during most conditions for the three participants. However, 5 of 7 of Alfred's sessions during the first High Disparity condition fell above the gray band, indicating a bias pattern for the left comparison. Nevertheless, accuracy remained relatively high during these sessions and, with a few exceptions (e.g., Sessions 12, 15, and 19),  $\log b$  (location) typically remained within the gray band for the remainder of the experiment. Therefore,  $\log b$  (location) across participants was minimal and/or transitory.

Figure 3 shows the obtained  $\log$  (base 10) reinforcer ratio between the two comparison stimuli (top row) and the obtained  $\log$  reinforcer ratio between the two comparison locations (bottom row) across conditions for all participants. A positive  $\log$  ratio indicates a greater number of reinforcers obtained in a session from Comparison Stimulus 1 than Comparison Stimulus 2 (top panel) and from the left-comparison location than the right-comparison location

(bottom panel). A negative log ratio indicates a greater number of reinforcers obtained in a session from Comparison Stimulus 2 than Comparison Stimulus 1 (top panel) and from the right-comparison location than the left-comparison location (bottom panel). Log values of  $\pm 1.0$  indicate a 10:1 or 1:10 arithmetic difference in reinforcer frequency between comparison stimuli or locations during a given session (typical of analyses used with the matching law; e.g., Baum, 2010).

The noteworthy patterns in the log reinforcer ratios from Figure 3 were the correspondences with  $\log b$  (stimulus) and  $\log b$  (location) from Figure 2. The clearest example is Harry's log reinforcer ratio between comparison stimuli in the top panel of Figure 3 approximating zero on average in the High Disparity conditions but was reliably negative in the Low Disparity condition – these log reinforcer ratios corresponded with patterns of  $\log b$  (stimulus) in Figure 2. Thus, bias increased for Comparison Stimulus 2 (Figure 2) along with a relative increase in reinforcer frequency for Comparison Stimulus 2 (Figure 3). With the Low Disparity condition decreasing discriminability, one interpretation is that biased responding toward Comparison Stimulus 2 was differentially reinforced, perhaps even in lieu of attending to the sample stimuli. In fact, Harry's  $\log d$  values were the lowest of the three participants during the Low Disparity condition. Increases in biases have been observed when discriminations become more difficult with increases in complexity of sample stimuli (e.g., Dube & McIlvane, 1997, 1999). Other examples of biases corresponding with log reinforcer ratios were less dramatic than Harry's performance but nevertheless apparent across all participants. A particularly noteworthy example is how Suzie's log reinforcer ratio between comparison stimuli trended negative toward the end of the Low Disparity condition when  $\log b$  (stimulus) also trended to negative values (see also Miles in Schneider et al., 2018). Continuation of the Low

Disparity condition might have produced a stimulus bias similar to Harry's. Overall, changes in log reinforcer ratios corresponded with shifts in bias measures, as shown in previous research (e.g., Alsop et al., 2016) and predicted by Davison and Tustin's (1978) framework. It is not possible, however, to specify whether biases drove changes in log reinforcer ratios or vice versa given the use of the rich ratio schedules. We discuss this issue in greater detail in the General Discussion.

Changes in percent correct were more closely related to changes in discriminability ( $\log d$ ) than biases ( $\log b$ ) for comparison stimuli or locations. For two participants (Alfred and Suzie), there was little relation between percent correct and  $\log b$  measures, although Suzie demonstrated a stimulus bias toward the end of the Low Disparity condition. For these participants, decreases in correct responding in the Low Disparity condition were generally associated with continued responding across comparison stimuli and locations in the array. Thus, reductions in correct responding were not related to a particular pattern of biased responding for Alfred and Suzie. In comparison, changes in Harry's percent correct was related to changes in  $\log b$  (stimulus). These findings reveal the utility of Davison and Tustin's (1978) analysis for quantitatively separating changes in conditional-discrimination performance by the specific error patterns comprising that performance, whether they are due to a single pattern (Alfred) or a combination of errors due to discriminability and bias (Suzie and Harry). These findings have implications for characterizing conditional-discrimination performance during clinical interventions – this quantitative framework could be used to identify error patterns from which procedures could be developed or employed specifically to target those error patterns.

Manipulating sample disparity primarily influenced discriminability for Alfred but produced a combination of discriminability and bias errors for Harry and briefly for Suzie. These

results are consistent with previous laboratory research that found both patterns of effects. For example, Gallagher and Alsop (2001) manipulated sample disparity with university students by adjusting the relative duration of auditory sample stimuli (tones) and observed decreases in  $\log d$  when decreasing sample disparity, while  $\log b$  did not change (see also Alsop, Rowley, & Fon, 1995; McCarthy & Davison, 1980). Findings showing only the influence of sample disparity on  $\log d$  and not  $\log b$  are consistent with the assumptions of Davison and Tustin's (1978) framework. However, not all findings support those assumptions. For example, Godfrey and Davison (1998) observed systematic changes in  $\log d$  with changes in sample disparity but also observed systematic changes in  $\log b$  with pigeons (see also Alsop & Davison, 1991; Nevin, Cate, & Alsop, 1993). It should be noted that the studies observing systematic changes in  $\log b$  with changes in sample disparity tended to manipulate sample disparity over a wider range of conditions than those that did not. Therefore, shifts in bias might have occurred more reliably in the present study with greater changes in sample disparity. To this effect, Harry's performance produced the lowest  $\log d$  values of all participants in the Low Disparity condition, and a reliable stimulus bias emerged in that condition.

### *Statistical Comparisons*

Linear mixed-effects modeling was performed using for each of the derived error metrics. For  $\log d$  values, modeling performed best with slopes and intercepts for Phase and Time varying at the individual subject level. Results indicated a significant effect for Phase ( $\beta = -0.86826$ ,  $SE = 0.084$ ,  $t = -10.43$ ,  $p < .001$ ), whereby the High disparity conditions had an overall value of 1.0368 and the Low disparity conditions an overall value of 0.16858. Modeling for each of the  $\log b$  values supported the varying of Phase and intercepts at the individual subject level. For  $\log b$  (stimulus), results indicated a non-significant effect for Phase ( $\beta = -0.26320$ ,  $SE = 0.249$ ,  $t = -$

1.054,  $p = .0368$ ) whereby the High disparity conditions had an overall  $\log b$  (stimulus) value of 0.06669 and the Low disparity conditions an overall value of -0.19651. Lastly, results for  $\log b$  (location) indicated a non-significant effect for Phase ( $\beta = 0.06588$ ,  $SE = 0.1013$ ,  $t = 0.650$ ,  $p = .555$ ) whereby the High disparity conditions had an overall  $\log b$  (location) value of 0.03877 and the Low disparity conditions an overall value of 0.10465. As such, these findings concur with those from visual analysis and support the conclusions that decreases in discriminability (i.e.,  $\log d$ ) were reliably predicted by the level of sample disparity.

## Experiment 2

In addition to manipulating sample disparity, laboratory research also examined the effect of manipulating disparity of the comparison stimuli during conditional discriminations (Alsop & Davison, 1991; Alsop, Rowley, & Fon, 1995; Gallagher & Alsop, 2001; Godfrey & Davison, 1998; Nevin, Cate, & Alsop, 1993). For example, Gallagher and Alsop manipulated disparity of comparison stimuli by adjusting the relative pixel density between stimuli. In these studies, decreasing comparison disparity decreased  $\log d$ , consistent with the effects of manipulating sample disparity described above. These studies suggest errors stemming from reductions in comparison-stimulus disparity also could be evaluated using Davison and Tustin's (1978) framework when teaching conditional discriminations. Therefore, Experiment 2 examined changes in disparity of the comparison stimuli on percent correct,  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location) in children diagnosed with ASD.

## Method

**Participants, setting, and materials.** Ari, Malik, and Pax participated in the second experiment. They were recruited and met the inclusion criteria from Experiment 1. Ari was 7 years old and had been receiving EIBI services for 38 months. An independent clinician not

associated with the center diagnosed him with ASD accompanied by language impairment and Avoidance/Restrictive Food Intake Disorder. His score was consistent with Level 3 on the VB-MAPP, with a score of 13 out of 15 on the Visual Performance/MTS section. Malik was four years old, diagnosed with ASD by an independent clinician not associated with the center, and had been receiving EIBI services for 12 months. His VB-MAPP score was consistent with Level 3, with a score of 15 on the Visual Performance/MTS section. Pax was four years old, diagnosed with ASD with language impairment by an independent clinician not associated with the center, and had been receiving EIBI services for 16 months. His VB-MAPP score also was consistent with Level 3, and a score of 10.5 out of 15 on the Visual Performance/MTS section.

All aspects of the setting and materials were consistent with those arranged in Experiment 1. Ari, Malik, and Pax were exposed to four, three, and four sessions, respectively, of training with the MTS procedure before moving to experimental sessions (data not shown but are available upon request). All three participants met the criterion in training that accuracy of independent responding was at or above 90% to begin experimental sessions.

## **Procedures**

All aspects of the procedures were consistent with those arranged in Experiment 1, with the exception that the sample stimuli remained identical across phases while comparison stimuli were manipulated across phases.

**Phases 1 and 3: High disparity.** The sample stimuli were pink and red, while the comparison stimuli were light pink and dark red (see Supporting Information). Correctly touching the light pink comparison after the pink sample and touching the dark red comparison after the red sample resulted in reinforcer delivery (edible and praise).

**Phase 2: Low disparity.** Comparison stimuli were more similar (lower disparity) than those arranged in Phase 1. Sample stimuli were identical with those in Phase 1. Comparison stimuli were the dark pink and light red samples. Correctly touching the dark pink comparison after the pink sample and touching the light red comparison after the red sample resulted in reinforcer delivery.

### **Procedural Integrity**

Procedural integrity in Experiment 2 was evaluated using the same methods described in Experiment 1 and was 100% for all three participants from 33% of sessions.

### **Analytical Plan**

Participant responding in Experiment 2 was statistically evaluated using the same manner as described in the methods of Experiment 1.

### **Results and Discussion**

Figure 4 shows percent correct,  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location) for Ari, Malik, and Pax across successive sessions of the High, Low, and High Disparity conditions. In the top row of Figure 4, percent correct was higher during the High Disparity conditions than during the Low Disparity conditions for all participants. Consistent with Experiment 1, Table 1 shows that reinforcers per session also were higher during the High Disparity conditions than during the Low Disparity conditions for all participants.

In the second row of Figure 4,  $\log d$  also was higher during the High Disparity conditions than during the Low Disparity conditions for all participants. Thus, decreasing comparison disparity generally reduced discriminability. In the third row,  $\log b$  (stimulus) did not change reliably across phases for any of the participants as a function of stimulus disparity, with the exception of a decrease in variability during the Low Disparity condition for Pax. In the bottom



row,  $\log b$  (location) also did not change reliably across phases for any of the participants as a function of stimulus disparity. However,  $\log b$  (location) gradually decreased toward the end of the Low Disparity condition for Pax and gradually increased at the beginning of the final High Disparity condition. The shift in  $\log b$  (location) suggests the decrease in comparison disparity likely produced a bias in addition to the decrease in discriminability. Nevertheless, the overall trend across participants is that decreases in percent correct were more related to changes in discriminability than shifts in stimulus or location biases.

Figure 5 shows the obtained log reinforcer ratio between the two comparison stimuli (top row) and between the two comparison locations (bottom row). As in Experiment 1, changes in bias tended to relate with changes in log reinforcer ratios. Specifically,  $\log b$  (stimulus) shown in Figure 4 changed with changes in log reinforcer ratio between comparison stimuli in Figure 5. Similarly, Figure 5 shows that  $\log b$  (location) corresponded with changes in log reinforcer ratio between comparison locations from Figure 4. For a particularly clear example with Pax, the log reinforcer ratio between comparison locations trended negative toward the end of the Low Disparity condition when  $\log b$  (location) also trended to negative values. Next, both the log reinforcer ratio between comparison locations and  $\log b$  (location) approached zero values upon returning to the High Disparity condition. These findings are similar to those observed with Suzie's log reinforcer ratios between comparison stimuli and  $\log b$  (stimulus) values at the end of the Low Disparity condition in Experiment 1. As with Experiment 1, changes in log reinforcer ratios might have functioned to shift biases, which has been shown in previous research (e.g., Alsop et al., 2016) and as predicted by Davison and Tustin (1978). However, it is not possible to determine precisely the direction of effects or interactions between biases and log reinforcer ratios. We discuss this issue in greater detail in the General Discussion.

Consistent with manipulating sample disparity in Experiment 1, manipulating comparison disparity across conditions in the present experiment produced corresponding changes in accuracy as measured by percent correct. Also similar to Experiment 1, changes in percent correct were closely related with changes in discriminability ( $\log d$ ), while biases ( $\log b$ ) for comparison stimuli or locations were not well related with changes in percent correct. Nevertheless, Pax's responding also showed a change in  $\log b$  (location) at the end of the Low Disparity condition. Previous laboratory research also has found complex relations between changes in comparison disparity and biases. For example, Gallagher and Alsop (2001) manipulated comparison disparity by adjusting the relative pixilation for two visual comparison stimuli. Similar to the present study, decreasing comparison disparity decreased  $\log d$  reliably but also tended to change  $\log b$  (see also Alsop & Davison, 1991; Godfrey & Davison, 1998; Nevin, Cate, & Alsop, 1993). It should be noted that the link between the present procedures and these previous studies is somewhat complicated by the fact that these studies also manipulated relative reinforcer frequencies between comparisons. Nevertheless, the present findings and those from previous studies clearly show that manipulating comparison disparity can produce complex error patterns involving both errors of discriminability and biases. These error patterns would not be elucidated using the traditional measure of percent correct and, as a result, demonstrate how Davison and Tustin's (1978) framework can describe different error patterns comprising conditional-discrimination performance. Thus, further development of these analyses for clinical use could provide analyses for identifying error patterns when teaching conditional discriminations.

### *Statistical Comparisons*

Linear mixed-effects modeling was performed using for each of the derived error metrics. For log  $d$  values, modeling performed best with slopes and intercepts for Phase and Time varying at the individual subject level. Results indicated a significant effect for Phase ( $\beta = -1.35193$ ,  $SE = 0.084$ ,  $t = -15.93$ ,  $p < .001$ ), whereby the High disparity conditions had an overall value of 1.46967 and the Low disparity conditions an overall value of 0.11774. Modeling for each of the log  $b$  values support the varying of individual intercepts alone at the individual subject level. For log  $b$ (stimulus), results indicated a non-significant effect for Phase ( $\beta = -0.12733$ ,  $SE = 0.058$ ,  $t = -2.181$ ,  $p = .032$ ) whereby the High disparity conditions had an overall log  $b$  (stimulus) value of 0.05145 and the Low disparity conditions an overall value of -0.07588. Lastly, results for log  $b$  (location) indicated a non-significant effect for Phase ( $\beta = -0.04266$ ,  $SE = 0.078$ ,  $t = -0.546$ ,  $p = .587$ ) whereby the High disparity conditions had an overall log  $b$  (location) value of -0.02623 and the Low disparity conditions an overall value of -0.06889. As such, these findings concur with those from visual analysis and support the conclusions that decreases in discriminability (i.e., log  $d$ ) were reliably predicted by the level of comparison disparity.

### General Discussion

The present experiments manipulated disparity of sample (Experiment 1) and comparison (Experiment 2) stimuli during conditional discriminations presented to children diagnosed with ASD. In both experiments, decreasing stimulus disparity decreased percent correct. We further analyzed the data from both experiments using Davison and Tustin's (1978) quantitative framework based on behavioral-choice and signal-detection theory. In doing so, we identified the error patterns comprising changes in conditional-discrimination performance across conditions, which is not possible with the more traditional measure of percent correct. The decreases in percent correct with reduced sample or comparison-stimuli disparity were more reliably due to

decreases in discriminability ( $\log d$ ), rather than changes in biases ( $\log b$ ) for a comparison stimulus or location. However, there were some isolated instances in which errors of both discriminability and bias contributed to decreases in percent correct, indicating the potential clinical usefulness of these quantitative methods for identifying error patterns. Specifically, these analyses could be used to (1) identify multiple sources of error patterns and (2) provide a basis for research that investigates antecedent- and consequence-based interventions based on those error patterns. Identifying interventions based on error patterns could lead to more individualized and, thereby, more efficacious interventions when teaching conditional discriminations (Kodak et al., 2011).

We observed reliable relations between percent correct and  $\log d$  when changing sample and comparison disparity between experiments. These findings suggest that decreases in sample and comparison disparity primarily impacted percent correct through changes in discriminability, rather than bias. If changes in discriminability ( $\log d$ ) exclusively accounted for changes in percent correct (rather than bias), these findings would support the suggestion of Davison and Tustin (1978) that the variables impacting discriminability (e.g., disparity) and bias (e.g., differential reinforcement) theoretically should be independent. There were some instances in which reduced sample disparity appeared to influence biases (see also Gallagher & Alsop, 2001). Any changes to bias with changes in sample or comparison disparity would indicate interactions between variables assumed by Davison and Tustin's framework to influence discriminability and bias independently, such as sample/comparison disparity and differential reinforcement (Alsop, 1991; Davison, 1991; Davison & Nevin, 1999). Therefore, changes in bias with decreases in comparison disparity would be inconsistent with predictions of Davison and Tustin's (1978) theoretical framework – variables influencing discriminability should be independent of changes

in bias, and vice versa. Additional research should assess a larger number of participants, as well as systematically manipulate variables predicted to influence biases (e.g., differences in reinforcer probabilities or amounts). Nevertheless, finding that these measures could identify changes both to discriminability and biases points to the usefulness of Davison and Tustin's descriptive framework. Specifically, Davison and Tustin's framework likely is less useful for *predicting* how particular variables will influence discriminability and biases but can be useful for *describing* error patterns produced by environmental changes (e.g., changes in sample or comparison disparity). As such, the present findings suggest Davison and Tustin's quantitative framework could be developed clinically to identify error patterns during the teaching of conditional discriminations.

When biases occur, it is important and potentially practical to ask what produced biases with changes to sample and comparison disparity in the present experiments. A simple explanation is differential reinforcement. Specifically, discriminations were difficult or impossible with low disparity and, as a result, reinforcement for solving the conditional discriminations was no longer reliably forthcoming. Choosing a particular comparison stimulus or location could at least reduce the effort required to obtain reinforcement on approximately 50% of trials, equivalent to reinforcement likelihood at chance performance with zero discriminability. Such differential reinforcement might be what underlies at least some instances of stimulus overselectivity often identified during complex discriminations in individuals diagnosed with cognitive disabilities (Dube & McIlvane, 1997, 1999). It should be noted that responding could be biased toward a particular comparison stimulus or location with low sample disparity, but low comparison disparity likely would only result in a bias for a comparison location because the comparisons are programmed to be difficult to discriminate. That is, the

comparisons would have to be discriminated even to show a bias to a particular stimulus, which is unlikely when comparison disparity is low.

With low sample or comparison disparity, participants also could begin choosing comparisons randomly to similar effect of reducing response effort to obtain reinforcement on approximately 50% of trials. Choosing randomly would appear only as sustained low discriminability and no change in bias. Thus,  $\log d$  would not distinguish random responding as a different pattern of choices from simply performing poorly while continuing to attend to the task.

The primary implication of the present findings and analyses for clinical application is the ability to identify and quantify error patterns in conditional-discrimination performance (Alsop & Davison, 1991; Godfrey & Davison, 1998; Nevin, Cate, & Alsop, 1993). Difficult conditional discriminations could produce changes in performance comprised of complex error patterns involving decreases in discriminability, shifts in bias, or a combination of bias and discrimination errors (Dube & McIlvane, 1997; Grow et al., 2011; Schneider et al., 2018; see Sidman, 1980). In the least, the present analyses allow researchers and, with further development, practitioners the ability to identify error patterns where the traditional measure of percent correct is insufficient. Further assessments could be developed to identify the processes underlying error patterns identified by this framework. Using these data, subsequent extensions of this methodology can be expanded to include other factors that may contribute to errors patterns, such as age, specific disability, and other co-morbid conditions.

With additional research aimed at applying these quantitative methods, a contribution of categorizing errors of discriminability and bias is the potential for implementing antecedent- and consequence-based interventions based on specific error patterns. Error patterns could yield data-based information for guiding practitioners in making clinical decisions when clients are not

making progress. For example, identifying whether a participant is making persistent errors because of biases or reduced discriminability can inform the practitioner or researcher to make either antecedent or consequent manipulations based on the kinds of errors emitted by the individual. If these equations depict that errors are due to low discriminability, practitioners could increase the salience of the sample stimuli to increase accuracy (e.g., elongate the samples; Fisher et al., 2019) or introduce differential observing responses (e.g., Dube & McIlvane, 1999). In contrast, location biases might suggest prompting and reinforcement for responses to other locations and/or repetition of error trials during error correction (e.g., Bourret, Iwata, Harper, & North, 2012).

The present analyses extend existing approaches to categorizing error patterns in conditional discriminations (Grow et al., 2011; Schneider et al., 2018). Grow et al. categorized errors as molar win-stay errors, molecular win-stay errors, and errors from position biases during conditional discriminations arranging auditory samples and visual comparisons. They arranged three comparison stimuli across nine trials per session with three children diagnosed with ASD. Win-stay errors are analogous to errors that would be categorized as  $\log b$  (stimulus) with the present Davison and Tustin (1978) analyses; position biases are analogous to errors categorized by  $\log b$  (location). All error types were calculated as a percentage of total trials within a session. The authors did not categorize errors based on discriminability. Their approach is limited by the data-collection process which is intensive and requires expertise to identify and analyze the potential error patterns. In contrast, recording errors based on chosen comparison stimuli and location within each session can be entered into a spreadsheet to calculate discriminability and bias errors, if an automated system is not used as in the present study. Data collection and analysis optimized for clinical settings is an area for further research with this general approach.

Relatedly, Schneider et al. (2018) also categorized error patterns during auditory-visual conditional discriminations as stimulus and position (location) bias among four different comparison stimuli using a touchscreen with four typically developing children. They set criteria for biases based on the number of choices for a given comparison out of 16 trials per session. Positive biases were defined as selecting a particular stimulus or location in six or more trials per session for three consecutive sessions. Negative biases were defined as selecting a particular stimulus or location in two or fewer trials per session for three consecutive sessions. They generally found biases to be mild and did not persist across more than three sessions. In addition, they concluded that biases detected in these analyses could not account for all errors. Therefore, these analyses only infer discriminability errors by those errors that are not from a source of bias (Grow et al., 2011). The advantage of using Davison and Tustin's analyses is that bias and discriminability error patterns are expressed quantitatively. Thus, biases are not categorized as present versus absent but as a matter of degree. Under some conditions, researchers or clinicians might choose to set criteria based on particular values of  $\log b$  (as we did with the gray bars indicating a single error in Figures 2 and 4), but such criteria are not necessary for assessing biases.

In the present experiments, we made stability judgments based only on percent correct. As a result, we observed changes in bias for some participants (Suzie and Pax) at the end of the Low Disparity conditions in both experiments. It would be useful to base stability on  $\log d$  and  $\log b$  values, too. We could have extended these conditions to examine whether those biases persisted. Clinically, observing the development of bias could prompt an intervention to reduce the likelihood that a persistent and disruptive error pattern develops. Moreover, tracking  $\log d$  and  $\log b$  values closely could suggest utility in assessing correlations of percent correct with  $\log$



$d$  and  $\log b$  values across sessions. Such correlations could be useful for assessing how such error patterns contribute to overall performance and making judgments about implementing antecedent- and consequence-based manipulations.

Our ultimate goal in conducting these experiments is to develop these analyses for use by clinicians. However, there are additional steps required before these analyses can be implemented effectively and practically in clinical situations. First, there were several differences between the conditional discriminations arranged in the present experiments and those typically arranged during clinical research and practice. Our procedures included only two comparison stimuli in contrast to arrays of three or more comparison stimuli in clinical research and practice. Davison and Tustin's (1978) analyses can accommodate more than two samples/comparisons by conducting the analyses in a pairwise fashion among the three or more trial types. For example, Godfrey and Davison (1998) reported that discriminability measures for pairs of stimuli did not change when additional sample and comparison stimuli were added to an array. As a result, there would be  $\log d$  and  $\log b$  values comparing performance individually between all combinations of trial types. Future research will need to address challenges of arranging multiple samples and comparisons, including conducting enough trials of each type to detect the levels of discriminability and bias among trial types and to prepare for possible interactions between different combinations of samples and comparisons.

Second, we arranged a touchscreen interface to allow for precise presentation of stimuli and data collection. Much clinical research and practice uses analog procedures with tangible or pictorial stimuli and manual data-recording methods (i.e., paper and pencil during tabletop instruction). As noted above, some approaches to characterizing biases are onerous and require expertise perhaps infrequently present in behavioral technicians (e.g., Grow et al., 2011).

Nevertheless, enhanced paper data-collection systems could be devised to record the placement of comparison stimuli and occurrence of responses across trials. With a carefully developed system, these analyses could be conducted in a spreadsheet after tallying the correct and error responses for all trial types.

We used sample and comparison stimuli that could be manipulated precisely to demonstrate the relation between disparity, percent correct, discriminability, and bias. Most clinical research and practice, however, arrange clinically relevant stimuli. Although this is a limitation for directly translating to application, this also points to a benefit of these analyses. Specifically, these analyses can quantify the qualitative differences between stimuli for individual participants or clients. Future research should examine the use of these analyses with more naturalistic visual sample and comparison stimuli, as well as other types of conditional discriminations (e.g., auditory-visual discriminations). However, a limitation of the Davison and Tustin (1978) model that could result in a limitation clinically is that  $\log d$  does not distinguish between changes in discriminability due to changes between sample stimuli versus changes in comparison stimuli. Nevertheless, we think the capacity to quantify discriminability and bias is a potential benefit making further exploration of these analyses worthwhile for clinical research and practice.

## References

- Alsop, B. (1991). Behavioral models of signal detection and detection models of choice. In M. L. Commons, J. A. Nevin, & M. C. Davison (Eds.), *Signal detection: Mechanisms, models, and applications* (pp. 39–55). Hillsdale, NJ: Erlbaum.
- Alsop, B. (2004). Signal-detection analyses of conditional discriminations and delayed matching-to-sample performance. *Journal of the Experimental Analysis of Behavior*, *82*, 57-69. doi: 10.1901/jeab.2004.82-57
- Alsop, B., Furukawa, E., Sowerby, P., Jensen, S., Moffat, C., & Tripp, G. (2016). Behavioral sensitivity to changing reinforcement contingencies in attention-deficit hyperactivity disorder. *The Journal of Child Psychology and Psychiatry*, *57*, 947-956.  
<https://doi.org/10.1111/jcpp.12561>
- Alsop, B., & Davison, M. (1991). Effects of varying stimulus disparity and the reinforcer ratio in concurrent-schedule and signal-detection procedures. *Journal of the Experimental Analysis of Behavior*, *56*, 67-80. <http://dx.doi.org/10.1901/jeab.1991.56-67>
- Alsop, B., & Rowley, R. (1996). Types of responding in a signal-detection task. *Journal of the Experimental Analysis of Behavior*, *65*, 561-574. doi: 10.1901/jeab.1996.65-561
- Alsop, B., Rowley, R., & Fon, C. (1995). Human symbolic matching-to-sample performance: Effects of reinforcer and sample-stimulus probabilities. *Journal of the Experimental Analysis of Behavior*, *63*, 53–70. doi:10.1901/jeab.1995.63-53
- Baum, W. M. (2010). Dynamics of choice: A tutorial. *Journal of the Experimental Analysis of Behavior*, *94*, 161–174. <https://doi.org/10.1901/jeab.2010.94-161>
- Beran, M. J., Menzel, C. R., Parrish, A. E., Perdue, B. M., Sayers, K., Smith, J. D., & Washburn, D. A. (2016). Primate cognition: Attention, episodic memory, prospective

- memory, self-control, and metacognition as examples of cognitive control in nonhuman primates. *WIREs Cognitive Science*, 7, 294–316. <http://dx.doi.org/10.1002/wcs.1397>
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 1-48. doi:10.18637/jss.v067.i01.
- Brown, G. S., & White, K. G. (2005) The optimal correction for estimating extreme discriminability. *Behavior Research Methods*, 37, 436-449.  
<https://doi.org/10.3758/BF03192712>
- Bourret, J. C., Iwata, B. A., Harper, J. M., & North, S. T. (2012). Elimination of position-biased responding in individuals with autism and intellectual disabilities. *Journal of Applied Behavior Analysis*, 45, 241-250. doi: 10.1901/jaba.2012.45-241.
- Carr, J., Nicolson, A., & Higbee, T. (2000). Evaluation of a brief multiple-stimulus preference assessment in a naturalistic context. *Journal of Applied Behavior Analysis*, 33, 353-357. doi: 10.1901/jaba.2000.33-353
- Cubicciotti, J. E., Vladescu, J. C., Reeve, K. F., Carroll, R. A., & Schnell, L. K. (2019). Effects of stimulus presentation order during auditory–visual conditional discrimination training for children with autism spectrum disorder. *Journal of Applied Behavior Analysis*, 52, 541–556. doi:10.1002/jaba.530
- Cumming, W., & Berryman, R. (1961). Some data on matching behavior in the pigeon. *Journal of the Experimental Analysis of Behavior*, 4, 281-284. doi:10.1901/jeab.1961.4-281
- Da Silva Barros, R., De Faria Galvao, O., & McIlvane, W. (2002) Generalized identity matching-to-sample in cebus apella. *The Psychological Record*, 52, 441-460.  
<https://doi.org/10.1007/BF03395197>

- Davison, M., & Jenkins, P. E. (1985). Stimulus discriminability, contingency discriminability, and schedule performance. *Animal Learning & Behavior*, *13*, 77-84.  
<http://dx.doi.org/10.3758/BF03213368>
- Davison, M., & McCarthy, D. (1987). The interaction of stimulus and reinforcer control in complex temporal discriminations. *Journal of the Experimental Analysis of Behavior*, *48*, 97-116. DOI: 10.1901/jeab.1987.48-97
- Davison, M., & Nevin, J. (1999). Stimuli, reinforcers, and behavior: an integration. *Journal of the Experimental Analysis of Behavior*, *71*, 439-482. doi:10.1901/jeab.1999.71-439
- Davison, M. C., & Tustin, R. D. (1978). The relation between the generalized matching law and signal-detection theory. *Journal of the Experimental Analysis of Behavior*, *29*(2), 331-336. doi:10.1901/jeab.1978.29-331
- DeHart, W. B., & Kaplan, B. A. (2019). Applying mixed-effects modeling to single-subject designs: An introduction. *Journal of the Experimental Analysis of Behavior*, *111*, 192-206. doi: 10.1002/jeab.507.
- Dube, W. V., & McIlvane, W. J. (1997). Reinforcer frequency and restricted stimulus control. *Journal of the Experimental Analysis of Behavior*, *68*, 303-316. <http://dx.doi.org.portal.lib.fit.edu/10.1901/jeab.1997.68-303>
- Dube, W. V., & McIlvane, W. J. (1999). Reduction of stimulus overselectivity with nonverbal differential observing responses. *Journal of Applied Behavior Analysis*, *32*, 25-33. doi:10.1901/jaba.1999.32-25
- Factari Software (2018). Paradigm stimulus presentation. Orlando, Florida. Retrieved from <http://www.paradigmexperiments.com/>

- Fisher, W. W., Kodak, T., & Moore, J. W. (2007). Embedding an identity-matching task within a prompting hierarchy to facilitate acquisition of conditional discriminations in children with autism. *Journal of Applied Behavior Analysis, 40*, 489–499. doi:10.1901/jaba.2007.40-489
- Fisher, W., Pawich, T., Dickes, N., Paden, A., & Toussaint, K. (2014). Increasing the saliency of behavior-consequence relations for children with autism who exhibit persistent errors. *Journal of Applied Behavior Analysis, 47*, 1-11. <https://doi.org/10.1002/jaba.172>
- Fisher, W. W., Retzlaff, B. J., Akers, J. F., DeSouza, A. A., Kaminski, A. J., & Machado, M. A. (2019). Establishing initial auditory-visual conditional discriminations and emergence of initial tacts in young children with autism spectrum disorder. *Journal of Applied Behavior Analysis, 52*, 1089-1106. <https://doi.org/10.1002/jaba.586>
- Gallagher, S., & Alsop, B. (2001). Effects of response disparity on stimulus and reinforcer control in human detection tasks. *Journal of the Experimental Analysis of Behavior, 75*, 183–203. doi:10.1901/jeab.2001.75-183
- Godfrey, R., & Davison, M. (1998). Effects Of Varying Sample- And Choice-stimulus Disparity On Symbolic Matching-to-sample Performance. *Journal of the Experimental Analysis of Behavior, 69*, 311–326. doi:10.1901/jeab.1998.69-311
- Green, G. (2001). Behavior Analytic Instruction for Learners with Autism: Advances in Stimulus Control Technology. *Focus on Autism and Other Developmental Disabilities, 16*, 72–85. <https://doi.org/10.1177/108835760101600203>.
- Grow, L. L., Carr, J. E., Kodak, T. M., Jostad, C. M., & Kisamore, A. N. (2011). A comparison of methods for teaching receptive labeling to children with autism spectrum

disorders. *Journal of applied behavior analysis*, 44, 475–498.

doi:10.1901/jaba.2011.44-475

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of  $d'$ . *Behavior Research Methods, Instruments & Computers*, 27, 46-51.

<http://dx.doi.org/10.3758/BF03203619>

Hutsell, B. & Banks, M. (2015). Effects of environmental and pharmacological manipulations on a novel delayed nonmatching-to-sample 'working memory' procedure in unrestrained rhesus monkeys. *Journal of Neuroscience Methods*, 251, 62-71.

doi:10.1016/j.jneumeth.2015.05.009

Hutsell, B. A., & Banks, M. L. (2017). Remifentanyl maintains lower initial delayed nonmatching-to-sample accuracy compared to food pellets in male rhesus monkeys. *Experimental and Clinical Psychopharmacology*, 25, 441–447.

doi:10.1037/pha0000154

Jones, B. M. & White, K.G. (1992). Sample-stimulus discriminability and sensitivity to reinforcement in delayed matching to sample. *Journal of the Experimental Analysis of Behavior*, 58, 159-172. doi: 10.1901/jeab.1992.58-159

Kangas, B. D., & Branch, M. N. (2008). Empirical validation of a procedure to correct position and stimulus biases in matching-to-sample. *Journal of the Experimental Analysis of Behavior*, 90, 103–112. doi:10.1901/jeab.2008.90-103

Kodak, T., Campbell, V., Bergmann, S., LeBlanc, B., Kurtz-Nelson, E., Cariveau, T., ... Mahon, J. (2016). Examination of efficacious, efficient, and socially valid error-correction procedures to teach sight words and preposition to children with autism spectrum disorder. *Journal of Applied Behavior Analysis*, 49, 532-547. doi:10.1002/jaba.310

- Kodak, T., Fisher, W. W., Clements, A., Paden, A. R., & Dickes, N. (2011) Functional assessment of instructional variables: Linking assessment and treatment. *Research in Autism Spectrum Disorders, 5*, 1059-1077. doi: 10.1016/j.rasd.2010.07.011
- Lovaas, O. I. (2003). Teaching individuals with developmental delays: Basic intervention techniques. Austin, TX, US: PRO-ED.
- MacDuff, G.S., Krantz, P.J., & McClannahan, L.E. (2001). Prompts and prompt- fading strategies for people with autism. In C. Maurice, G. Green, & R. M. Foxx (Eds.), *Making a difference: Behavioral intervention for autism* (pp. 37-50). Austin, TX: Pro-Ed.  
Retrieved from: [psycnet.apa.org](http://psycnet.apa.org)
- McCarthy, D. C. (1991). Behavioral detection theory: Some implications for applied human research In *Signal Detection: Mechanisms, Models, and Applications* (pp. 239-255). Psychology Press. Retrieved from <http://books.google.com>
- McCarthy, D. & Davison, M. (1980). Independence of sensitivity to relative reinforcement rate and discriminability in signal detection. *Journal of the Experimental Analysis of Behavior, 34*, 273-284. DOI:10.1901/jeab.1980.34-273
- McGhan, A., & Lerman, D. (2013). An assessment of error-correction procedures for learners with autism. *Journal of Applied Behavior Analysis, 46*, 626-639.  
<https://doi.org/10.1002/jaba.65>
- McIlvane, W. J., Kledaras, J. B., Gerard, C. J., Wilde, L., & Smelson, D. (2018). Algorithmic analysis of relational learning processes in instructional technology: Some implications for basic, translational, and applied research. *Behavioural processes, 152*, 18–25.  
doi:10.1016/j.beproc.2018.03.001



- Nall, R. W., Rung, J. M., & Shahan, T. A. (2019). Resurgence of a target behavior suppressed by a combination of punishment and alternative reinforcement. *Behavioural Processes*, *162*, 177-183. <https://doi.org/10.1016/j.beproc.2019.03.004>
- Nevin, J. A., Cate, H., & Alsop, B. (1993). Effects of differences between stimuli, responses, and reinforcer rates on conditional discrimination performance. *Journal of the Experimental Analysis of Behavior*, *59*, 147–161. doi:10.1901/jeab.1993.59-147
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Saunders, K. J., & Spradlin, J. E. (1989). Conditional discrimination in mentally retarded adults: the effect of training the component simple discriminations. *Journal of the Experimental Analysis of Behavior*, *52*, 1–12. doi:10.1901/jeab.1989.52-1
- Schneider, K.A., Devine, B., Aguilar, G., & Petursdottir, A. (2018). Stimulus presentation order in receptive identification tasks: A systematic replication. *Journal of Applied Behavior Analysis*, *51*(3), 634-646. DOI: 10.1002/jaba.459
- Shahan, T. A., & Podlesnik, C. A. (2006). Divided attention performance and the matching law. *Learning & Behavior*, *34*, 255-261. <http://dx.doi.org/10.3758/BF03192881>
- Shahan, T. A., & Podlesnik, C. A. (2007). Divided attention and the matching law: Sample duration affects sensitivity to reinforcement allocation. *Learning & Behavior*, *35*, 141-148. <http://dx.doi.org/10.3758/BF03193049>
- Sidman, M. (1960). *In Tactics of scientific research: Evaluating experimental data in psychology*. Oxford, England: Basic Books.
- Sidman, M. (1980). A note on the measurement of conditional discrimination. *Journal of the Experimental Analysis of Behavior*, *33*, 285–289.

<https://doi.org/10.1901/jeab.1980.33-285>.





- Stromer, R., McIlvane, W.J., Dube, W.V., & Mackay, H.A. (1993). Assessing control by elements of complex stimuli in delayed matching to sample. *Journal of Experimental Analysis of Behavior, 59*, 83-102. DOI: 10.1901/jeab.1993.59-83
- Sundberg, M. L. (2008). VB-MAPP: *Verbal behavior milestones assessment and placement program*. Concord, CA: AVB Press.
- Tiger, J. H., Hanley, G. P., & Bruzek, J. (2008). Functional communication training: a review and practical guide. *Behavior Analysis in Practice, 1*, 16–23. doi:10.1007/BF03391716
- Truppa, V., Mortari, E. P., Garofoli, D., Privitera, S., & Visalberghi, E. (2011). Same/different concept learning by capuchin monkeys in matching-to-sample tasks. *PLoS ONE, 6*. Article ID e23809. <http://dx.doi.org/10.1371/journal.pone.0023809>
- Williams, D. C., Johnston, M. D., & Saunders, K. J. (2006). Intertrial sources of stimulus control and delayed matching-to-sample performance in humans. *Journal of the Experimental Analysis of Behavior, 86*, 253–267. <https://doi.org/10.1901/jeab.2006.67-01>
- Zentall, S. (2005). Theory- and evidenced-based strategies for children with attentional problems. *Psychology in the Schools, 42*, 821-836. <https://doi.org/10.1002/pits.20114>

Table 1. Mean and range reinforcer frequency across 24-trial sessions of High, Low, and High Disparity conditions for all participants of Experiments 1 and 2.

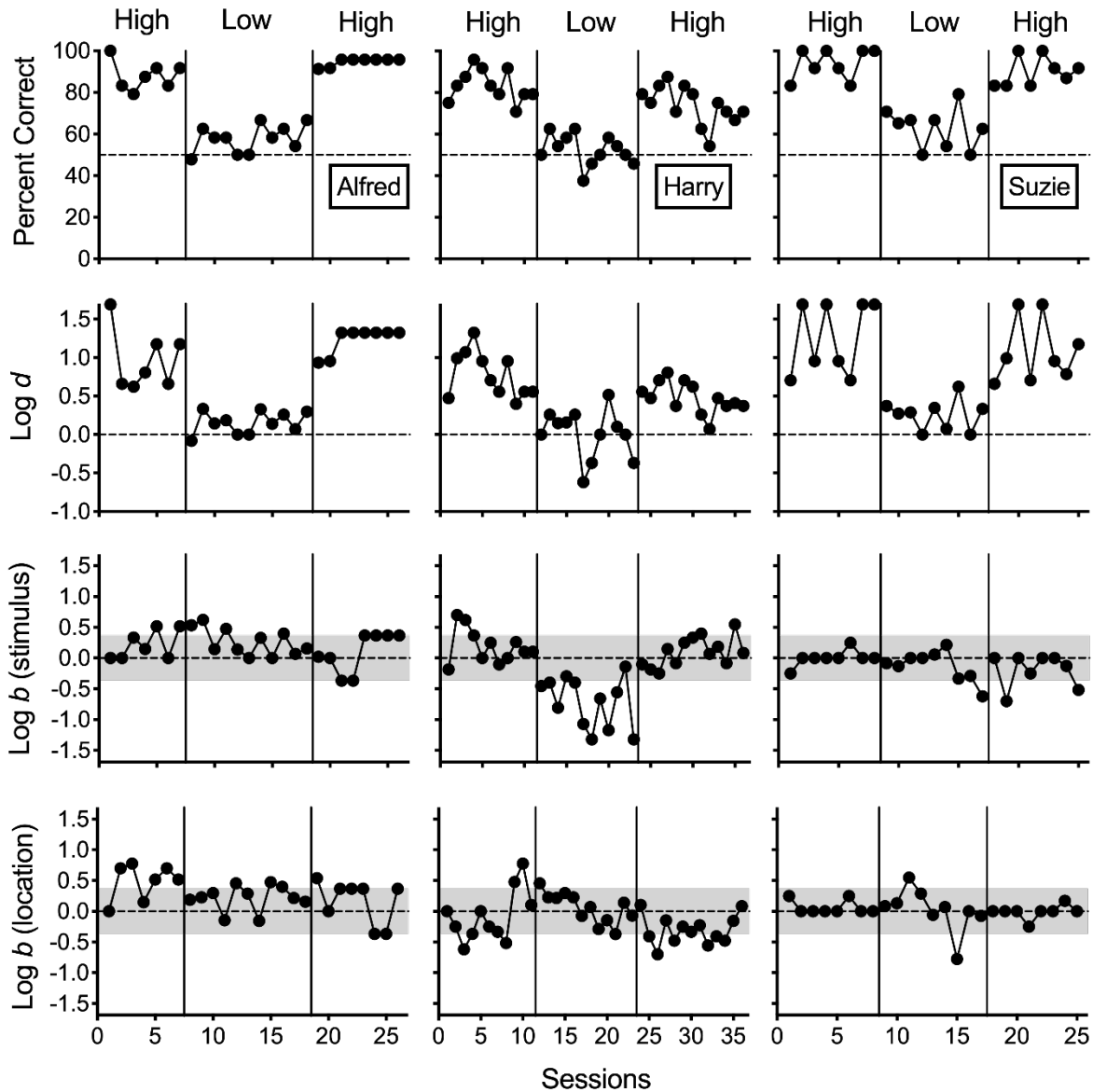
Experiment	Participant	Condition	Mean	Range	
				Min	Max
1	Alfred	High	21	19	24
		Low	14	12	16
		High	23	21	23
	Harry	High	20	17	23
		Low	13	9	15
		High	18	13	21
	Suzie	High	23	20	24
		Low	15	12	19
		High	22	20	24
2	Ari	High	23	22	24
		Low	13	9	23
		High	23	20	24
	Malik	High	24	22	24
		Low	13	9	17
		High	24	23	24
	Pax	High	23	19	24
		Low	12	8	15
		High	21	16	23

		Comparisons	
		$C_1$	$C_2$
Samples	$S_1$	11	12
	$S_2$	21	22

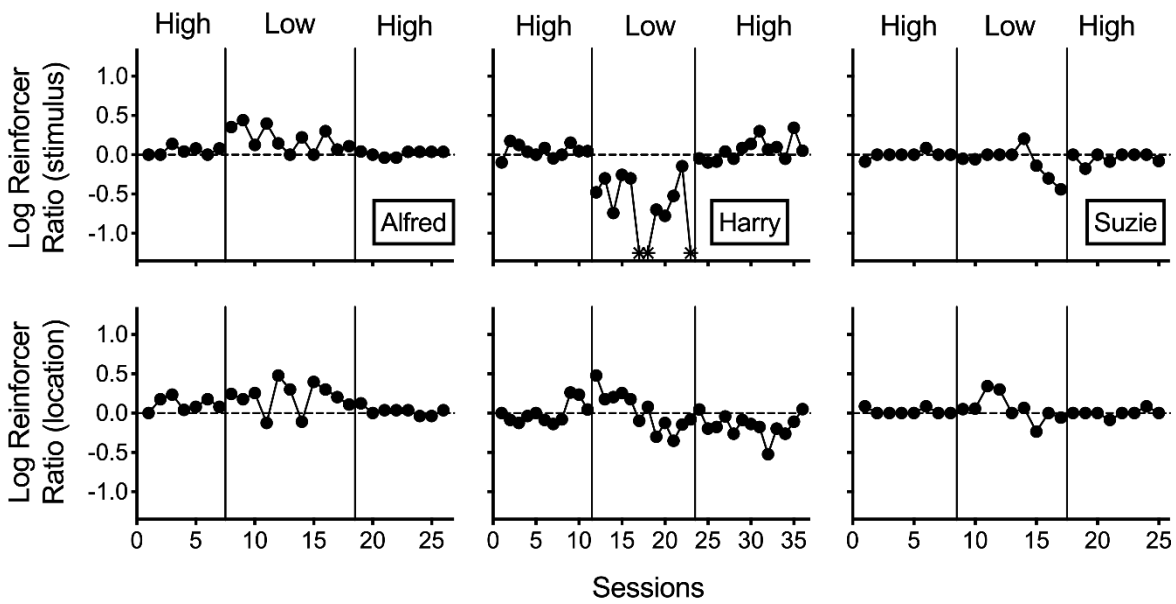
  

Dog	
	
<i>Correct</i> <sub>11</sub>	<i>Error</i> <sub>12</sub>
Cat	
	
<i>Error</i> <sub>21</sub>	<i>Correct</i> <sub>22</sub>

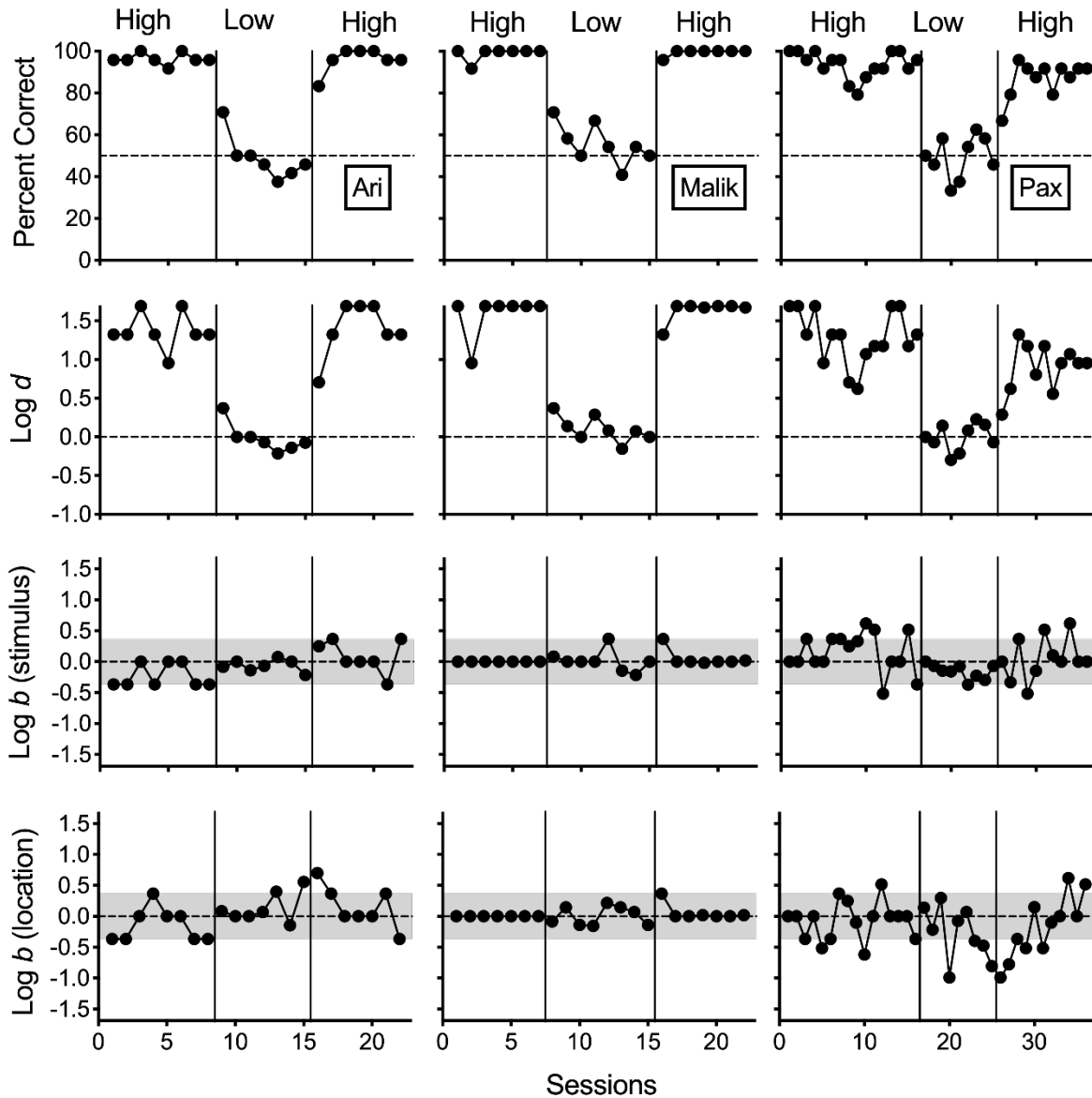
**Figure 1.** The top panel shows a conditional-discrimination matrix. The matrix shows the two samples ( $S_1$  and  $S_2$ ) and responses to the two comparisons ( $C_1$  and  $C_2$ ), with correct responses (11, 22) and errors (12, 21). The bottom panel shows a hypothetical example of a conditional discrimination if the written word Dog were  $S_1$  and the written word Cat were  $S_2$ .



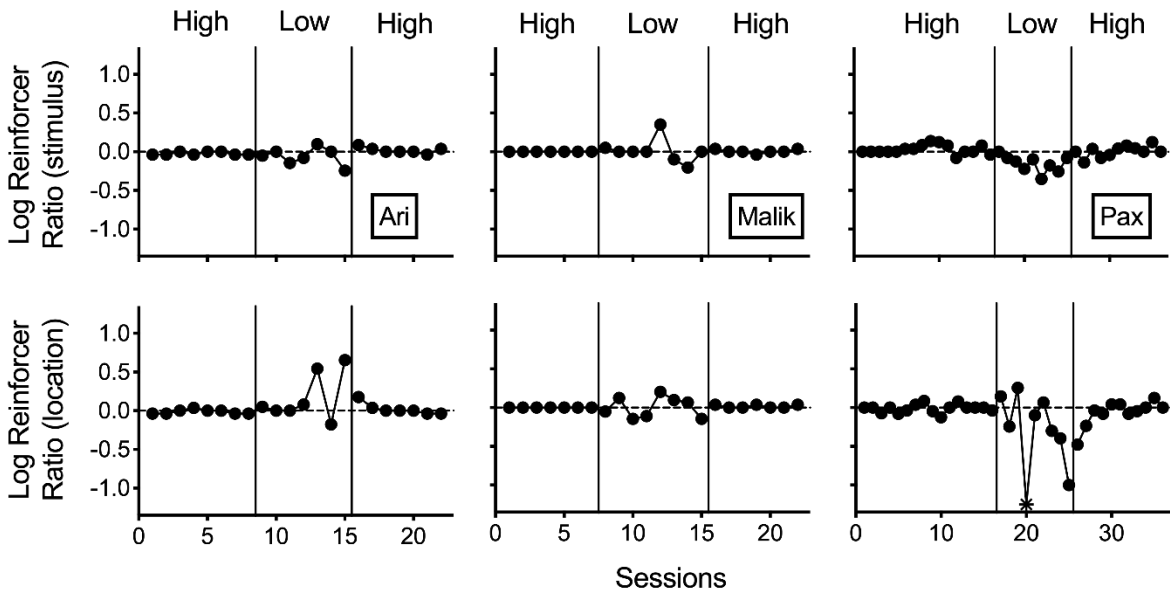
**Figure 2.** Percent correct,  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location) across sessions of the High Disparity and Low Disparity conditions for Alfred (left column), Harry (middle column) and Suzie (right column) in Experiment 1. Dashed lines indicate chance performance for percent correct, zero discriminability for  $\log d$ , and zero bias for  $\log b$  (stimulus) and  $\log b$  (location). The gray bands in the bottom two panels range from  $\pm 0.368$  and indicates the values of  $\log b$  obtained with only a single error out of 24 trials.



**Figure 3.** Log reinforcer ratio (stimulus) and log reinforcer ratio (location) across sessions of the High Disparity and Low Disparity conditions for Alfred (left column), Harry (middle column) and Suzie (right column) in Experiment 1. Dashed lines indicate equal obtained reinforcer frequencies between the two comparison stimuli (top panel) and locations (bottom panel). Data points that are asterisks are when zero reinforcers were earned from one comparison stimulus or location during that session and are placed at -1.5.



**Figure 4.** Percent correct,  $\log d$ ,  $\log b$  (stimulus), and  $\log b$  (location) across sessions of the High Disparity and Low Disparity conditions for Ari (left column), Malik (middle column) and Pax (right column) in Experiment 2. Dashed lines indicate chance performance for percent correct, zero discriminability for  $\log d$ , and zero bias for  $\log b$  (stimulus) and  $\log b$  (location). The gray bands in the bottom two panels range from  $\pm 0.368$  and indicates the values of  $\log b$  obtained with only a single error out of 24 trials.



**Figure 5.** Log reinforcer ratio (stimulus), log reinforcer ratio (location), and reinforcer frequency across sessions of the High Disparity and Low Disparity conditions for Ari (left column), Malik (middle column) and Pax (right column) in Experiment 2. Dashed lines indicate equal obtained reinforcer frequencies between the two comparison stimuli (top panel) and locations (bottom panel). Data points that are asterisks are when zero reinforcers were earned from one comparison stimulus or location during that session and are placed at -1.5.