Psychol Rec (2015) 65:49–65
DOI 10.1007/s40732-014-0087-y

A|B|A|I Association for Behavior Analysis International

ORIGINAL ARTICLE

An Examination of the Effects of Feedback Accuracy on Academic Task Acquisition in Analogue Settings

Jason M. Hirst · Florence D. DiGennaro Reed

Published online: 26 July 2014

Abstract Performance feedback is a common procedure used in a variety of settings to change behavior. Although reviews of the literature have identified a number of characteristics of performance feedback that are predictive of effectiveness, little research has examined the influence of feedback *accuracy* on behavior. The purpose of the present study was to examine both the short-term and long-term effects of inaccurate feedback on the acquisition of match-to-sample tasks. The first study adopted a translational, human operant paradigm to evaluate the effects under highly controlled conditions. Undergraduate students were presented an arbitrary match-to-sample task on a computer. Feedback accuracy was manipulated in an initial phase followed by a condition where only accurate feedback was provided. The second study extended these findings to a more applied setting and population. The results of both studies demonstrated that exposure to inaccurate feedback generally resulted in the failure to acquire the tasks. Of even greater importance, for most participants a carryover effect was obtained, represented by a delay to acquisition following the improvement of feedback accuracy. The patterns of responding obtained suggested both a reinforcement function and rule-governance. Further inquiries into the function of feedback may facilitate an interpretation of feedback accuracy within the context of procedural fidelity. These findings may have implications for organizational and educational settings where circumstances lead individuals to be exposed to conflicting sources of feedback.

Keywords Feedback accuracy · Performance feedback · Instruction · Match-to-sample · Rule-governed behavior

J. M. Hirst · F. D. DiGennaro Reed (✉)
Department of Applied Behavioral Science, University of Kansas, 4056 Dole Human Development Center, 1000 Sunnyside Avenue, Lawrence, KS 66045-7555, USA
e-mail: fdreed@ku.edu

Performance feedback involves the delivery of information to an individual regarding past performance and indicates how an individual can improve his or her performance in the future (Daniels 1994; Prue and Fairbank 1981). It is one of the most commonly used procedures to change behavior in applied settings. For example, within the field of organizational behavior management, more than half of the reports published in the *Journal of Organizational Behavior Management* include some form of feedback in their methodology (Alvero et al. 2001). Feedback also plays an important role in education where student performance is likely influenced, at least in part, by feedback from teachers in the form of grades, verbal praise, or written corrections on assignments.

Despite its use in practice, the basic and translational research on the necessary and sufficient characteristics of feedback is surprisingly sparse. For example, we were able to find only a small number of laboratory analyses of the characteristics of feedback, including frequency (e.g., Kang et al. 2003) and immediacy (e.g., Mason and Redmon 1992). The applied literature has presented some evidence for the characteristics of performance feedback through meta-analyses. Notably, Balcazar et al. (1985) proposed that feedback efficacy varies along the dimensions of source, format, frequency, content, media, and level of privacy. Others have proposed similar characteristics and demonstrated varying efficacy associated with particular characteristics. (e.g., Alvero et al. 2001). A conceptual paper by Peláez and Moreno (1998) suggested that rules vary in accuracy—the degree to which they accurately describe underlying contingencies. An extension of the concept of accuracy might also be applied to feedback. Although the notion that feedback should be accurate might be implied, the literature has shown that many procedures in behavior analysis are not implemented in a consistent and accurate manner, resulting in treatment failures (e.g., Sanetti and Kratochwill 2009). By extension, errors in feedback might also influence the efficacy and outcomes of interventions.

However, few studies have experimentally evaluated feedback accuracy.

When intervention procedures are implemented by individuals with less training, deviations from prescribed procedures are more likely to occur. This phenomenon has been shown when less rigorous training procedures are used to prepare educators to implement behavioral interventions (e.g., Ducharme and Feldman 1992; Rose and Church 1998). Deviations from procedures may also be observed in other contexts. For example, a common practice in educational settings is to rely on peers to deliver intervention or instruction (e.g., a teacher organizes dyads or cooperative learning groups wherein students instruct their peers). Although a core benefit to peer-mediated intervention is its resource efficiency, a review of the training procedures used in published research revealed that they include an ongoing monitoring component to ensure procedural fidelity, or accurate implementation of the intervention (Chan et al. 2009). In these studies, monitoring involved continuous observation, corrective feedback, and redirection when peer interventionists deviated from prescribed procedures. These monitoring techniques are resource intensive, which detracts from the benefit of using peer-mediated intervention and may not occur outside of research settings. In its absence, procedural fidelity may decline and, thus, disrupt learning.

Similar concerns can be found in the literature on organizational behavior management. In a case study of a furniture distribution company, Mihalic and Ludwig (2009) depicted how the implementation of a flawed measurement system for recording employee errors did not accurately or consistently identify errors. This led to a failure of the system to deliver feedback to employees that errors were made, incorrectly indicating that their performance was correct. The authors reported that the failure of the measurement system resulted in a subsequent failure of the incentive system. Unreported employee errors led to the delivery of incentives following errors. Thus, for more than 5 years, employees received inaccurate feedback and monetary reinforcement for making errors, which hindered improvement efforts. When an employee faces competing contingencies in the workplace, contradictory feedback from opposing parties may also influence performance. For example, Cooper (2006) described competition between safety initiatives and costs and misperceptions about the intervention process (i.e., releasing workers from job responsibilities to observe peers, the perception that peer observation is akin to spying). It would be possible that despite ostensibly committing to a safety initiative, discrepant sources of feedback support productivity over safety. Myers et al. (2010) described a system in place at a refinery in which incentives were delivered contingent upon the absence of reportable incidents rather than the presence of important safety behaviors, possibly disincentivizing the accurate reporting of data and thereby producing inaccuracies in the feedback delivered to employees. The above examples provide a foundation for the necessity of understanding how feedback reliably influences behavior under conditions in which inaccurate or opposing feedback is delivered. Further research and modeling of applied concerns is necessary to identify the breadth and scope of application of accuracy as a dimension of feedback.

In a preliminary effort to evaluate how feedback accuracy might influence behavior, Hirst et al. (2013) evaluated the effects of feedback accuracy during the acquisition of an arbitrary match-to-sample task. Participants received inaccurate feedback (i.e., "Correct" following incorrect responses or "Incorrect" following correct responses) during a proportion of trials. The degree of accuracy of the feedback was determined by the proportion of trials during which inaccurate versus accurate feedback was delivered: 25 %, 50 %, 75 %, or 100 % accuracy. Participants in the 100 % accuracy group, who were exposed only to accurate feedback following responses to the task, mastered the task and rapidly achieved the criterion of 15 consecutive correct responses. However, participants exposed to inaccurate feedback (i.e., 25 %, 50 %, and 75 % accuracy) failed to meet this criterion despite occasionally emitting the correct response and receiving accurate feedback. That is, correct responding was hindered by exposure to feedback specifying incorrect contingencies on previous or subsequent trials. In a second condition, participants who were exposed to inaccurate feedback were given 100 % accurate feedback for the remainder of the study. Prior exposure to inaccurate feedback appeared to have a persistent, negative effect, even after receiving consistent and accurate feedback. Some participants mastered the task only after a substantial delay, and others failed to master the task at all.

Taken together, these findings suggest that feedback errors (i.e., inaccurate or inconsistent feedback) may partially account for the reduced effectiveness of behavioral interventions. The goal of the present study was to extend the extant literature on feedback through a systematic replication and extension of Hirst et al. (2013). We adopted a paradigm of translational research described by Stokes (1997) as "use-inspired basic research" (p. 74), in which understanding of fundamental principles is pursued but with explicit considerations for practical uses (see also Mace and Critchfield 2010). In the present study, a phenomenon of interest to practitioners—feedback accuracy—was brought into the laboratory for examination to better understand the principles behind it and to improve practice. Study 1 adopted a highly controlled, human operant paradigm to examine the effects of feedback accuracy with undergraduate students in laboratory conditions using an arbitrary task. Study 2 assessed the generality of findings to a less controlled, analogue educational setting and to a more applied population (i.e., preschool-age children).

Study 1

Method

Participants and Setting The participants for Study 1 were 64 undergraduate students (51 females, 13 males) ages 18 to 53 years (M=20.06) who were enrolled in introductory-level courses in applied behavior analysis at a large Midwestern university. A post hoc exclusionary criterion was applied for participants whose average response latency was less than 0.5 s during the initial condition. We applied this criterion as a means to estimate whether participants were sufficiently attending to the procedures and stimuli of the study. The criterion was based on data from a previous study (Hirst et al. 2013) and resulted in the exclusion of five participants. In addition, two participants were excluded because they did not complete 500 trials within the 1-hr time limit. Sessions for the present study took place in a large computer lab measuring 9 m by 6 m, containing 20 computers.

Materials The task was an arbitrary match-to-sample task for which five nonsense shapes and nine nonsense names were developed. The five shapes were each assigned a nonsense name containing two syllables, each beginning with a different phoneme. Four additional names were created to serve as distracters. The five target shapes and corresponding nonsense names are provided in Table 1.

The match-to-sample task was presented by a computer program written in Microsoft® Visual Basic .Net. The program interface was designed to occupy the entire screen, making other features of the computer inaccessible. Each trial consisted of the presentation of a sample stimulus and comparison stimuli, a participant response, and a feedback period lasting 3 s. Participants were presented with one nonsense shape in black as a sample stimulus in a white rectangle measuring 6.5 cm by 6.5 cm. The comparison stimuli for each trial were five nonsense names presented on a vertical column of buttons 2.0 cm to the right of the sample stimulus. The array consisted of three target names (i.e., the correct answer and two additional target names associated with other shapes) and two distracter names. The comparison stimuli presented on each trial were selected using a random algorithm built into the programming language. In addition, the order of the names in the comparison array was randomized for each trial using the same random function. For reference, screenshots of the array of stimuli are presented in Hirst et al. (2013). The sequence of shapes presented across trials was determined a priori using a random number generator found at www.random.org. Constraints were placed on the randomness of the sequence to ensure that all five shapes were presented an equal number of times during each condition and were never presented more than twice consecutively. The program was designed to end after the participant had completed a total of 500 trials across two conditions, or after an elapsed time of 1 hr.

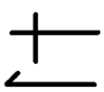
Dependent Variables and Response Measurement The dependent variables were participant responses to the task and the latency in trials until the mastery criterion was met. Correct responses were defined as selecting the nonsense name associated with the shape presented during the trial. Incorrect responses included selecting a target name associated with a shape not being displayed or a distracter name. The mastery criterion for the present study was set at 15 consecutive correct responses.

Experimental Design and Procedure A between-groups design was used to compare four levels of feedback accuracy. Participants were first randomly assigned to one of four groups, each containing 16 participants: 25 %, 50 %, 75 %, or 100 % accuracy (the latter served as a comparison group). Groups were further divided into four subgroups (n=4), which determined the duration of exposure to the first condition (inaccurate feedback). Durations examined in the present study were 160, 200, 260, or 300 trials. The selection of condition durations was informed by the data presented in Hirst et al. (2013). Specifically, the minimum number of trials was set at 160, based on the average number of trials to criterion of the comparison group in that study. Other values were chosen arbitrarily in order to test a wide range of durations. However, the values were also constrained to multiples of 20 in order to ensure that each of the five shapes was presented an equal number of times and the obtained level of accuracy matched the programmed level.

Inaccurate Feedback Condition During the inaccurate feedback (IF) condition, the program delivered IF following a designated proportion of participant responses. Inaccurate feedback was defined as either presenting the feedback "Incorrect" following a correct response or the feedback "Correct" following an incorrect response. Both forms of feedback were displayed for 2 s below the sample stimulus following a participant response. "Correct" was displayed in black, bold font in a green rectangle measuring 6.6 cm by 1.5 cm. "Incorrect" was displayed similarly in a red rectangle of the same dimensions. The proportion of these two types of errors was not held constant, as the type of feedback error was dependent on participant responses. These procedures resemble those of Study 1 in St. Peter Pipkin et al. (2010) wherein participants were exposed to combined errors of commission (i.e., reinforcing problem behavior) and errors of omission (i.e., failing to reinforce appropriate behavior).

Depending on the group to which participants were assigned, the proportion of trials during which the program delivered inaccurate feedback varied. The sequence of

Table 1 Target Shapes and Corresponding Names

Name	Zitaaf	Raopol	Smuzy	Punfi	Bifdo
Shape					

inaccurate and accurate feedback trials was determined by a pseudorandom sequence, constrained to ensure the ratio of inaccurate and accurate feedback was equal across all five shapes.

Accurate Feedback Condition During this condition, feedback errors were no longer committed. All feedback delivered by the program following responses was accurate ("Correct" after a correct response and "Incorrect" after an incorrect response). No change in the program interface occurred concurrently with the change in condition. Participants continued the task until a total of 500 trials had been completed between both the IF and accurate feedback (AF) conditions. In addition, the participants assigned to the 100 % accuracy comparison group were never exposed to IF and completed 500 trials in the AF condition.

Data Analysis In addition to visual inspection of cumulative records of correct responses, between-group analyses were conducted to compare task acquisition across the four levels of feedback accuracy. To standardize rates of acquisition across participants and groups, an index was generated using the area under the curve (AUC) of the cumulative record. AUC is a relatively common method for comparing curvilinear data across subjects, particularly in behavioral economics (see Myerson et al. 2001). AUC was calculated using the trapezoidal method over other methods because it does not require curve fitting or integral calculus but also provides a relatively accurate estimate of area. The formula used was:

$$AUC = \sum (x_2 - x_1)\left[\frac{y_1 + y_2}{2}\right]$$

This value was calculated per participant and per phase. Because the duration in trials of each phase varied across participants, AUC was converted to a standardized proportion by also calculating the maximum AUC that would be obtained if participants emitted perfect responding to the task for each phase (i.e., the AUC of a line with a slope of 1) and dividing the AUC of the cumulative record by this value, yielding a percentage (%AUC).

Between-group comparisons were made using two nonparametric statistical analyses (Kruskal-Wallis ANOVA and multiple comparison posttest) to determine whether rates of acquisition were differentially affected by varying levels of feedback accuracy in either phase (i.e., both in the presence of IF and due to prior exposure to IF).

Finally, conditional probabilities were calculated as $P(R|S)$, in which R is a specific response (nonsense name) and S is the sample shape presented. Probabilities were calculated for each possible response (five target names and four distracters) in the presence of each of the five sample stimuli. This analysis was used to evaluate whether providing inaccurate feedback resulted in the strengthening of a specific incorrect response (one response with an elevated conditional probability) or resulted in a persistence of random responding (approximately equal probabilities of all responses). Because five responses were available during each trial, a probability of 0.2 would suggest that the response was not emitted above chance levels whereas a probability substantially higher than .2 would suggest that the response was strengthened in the presence of inaccurate feedback.

Results and Discussion

Cumulative records for participants in the 25 % accuracy group are presented in Fig. 1. In addition to cumulative records of correct responses, a gray region superimposed on each graph represents the range of acquisition rates for participants in the 100 % accuracy group for comparison. Acquisition for participants in the 25 % accuracy group was generally low, indicated by the shallow slope of the cumulative records to the left of the phase line. There was very little overlap with the data from participants in the comparison group during the initial learning curve. None of the participants met the mastery criterion during IF. After the accuracy of feedback improved, acquisition of the task varied across participants. Some participants acquired the task rapidly following the removal of errors, indicated by an inflection point in the cumulative record shortly after the phase line (102, 104, 202, 301). Other participants acquired the task after a longer exposure to improved instruction indicated by an inflection point further to the right of the phase line (103, 302, 303, 304, 403, 404). Two of the 16 participants in this group did not acquire the task prior to the session ending at 500 trials (201, 204). Of the 14 participants who did master the task, the average number of trials following the condition change at which criterion was met was 136.43 (range: 48–220).

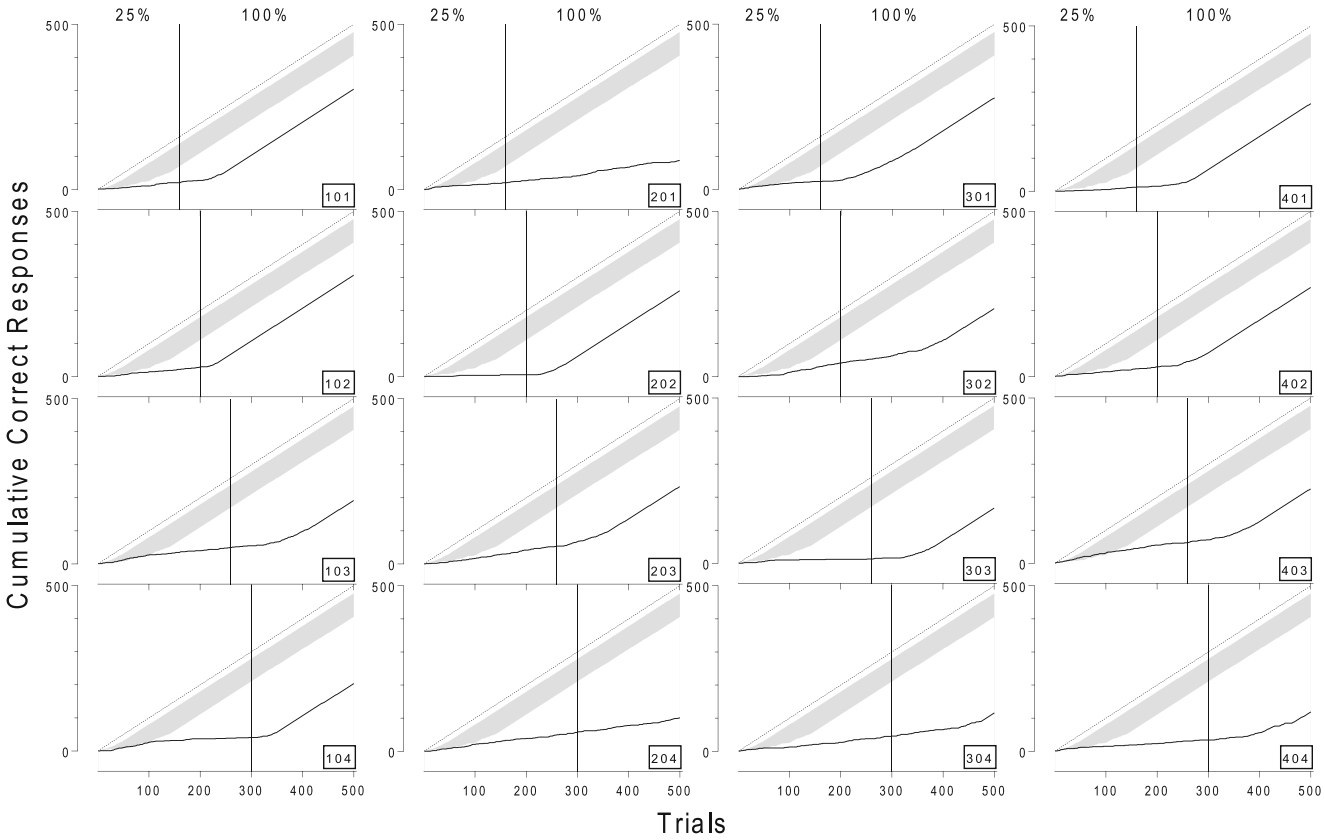


Fig. 1 Cumulative records for participants in the 25 % accuracy group

Figure 2 displays data from participants assigned to the 50 % accuracy condition. Acquisition was low in the presence of IF. Two participants (208, 407) showed some overlap initially with the comparison group, but prior to the end of IF, performances differentiated. The other 14 participants' performances were clearly differentiated from the comparison group. Also, acquisition varied similarly following the removal of feedback errors with some participants rapidly acquiring the task, while some participants acquired the task more slowly. Four out of 16 participants did not meet the mastery criterion (306, 307, 308, 408). Of the 12 participants who met the criterion, the average number of trials relative to the phase change at which criterion was met was 101.75 (range: 61–139).

Data for participants assigned to the 75 % accuracy group are presented in Fig. 3. The data obtained from this group differed from participants in the 25 % and 50 % accuracy groups. Acquisition in the presence of IF was higher as indicated by the steeper slopes to the left of the phase line. Four of 16 participants met the mastery criterion prior to the phase change (209, 309, 312, 410) during IF. Additionally, acquisition for these participants fell within the range of the comparison group, suggesting that exposure to feedback errors during only 25 % of trials in IF did not negatively influence learning. All but two participants quickly acquired the task following the removal of errors. However, two participants did not acquire the task (111, 112). Of the 10 participants who acquired the task only after errors were removed, the average number of trials in AF to criterion was 54 (range: 14–88).

Participants in the 25 % and 50 % accuracy groups did not acquire the task during IF. Acquisition occurred only after feedback accuracy improved. In addition, a delay to the improvement in acquisition was obtained suggesting that the negative influence of poor instruction on learning persisted under AF. Participants exposed to 75 % accuracy showed higher rates of acquisition during IF with four participants mastering the task despite the errors.

The number of trials in AF before the mastery criterion was met for all participants are plotted in Fig. 4, with each data point representing a single participant. Data points falling within the gray region represent mastery prior to the removal of feedback errors. Comparing the medians of the three groups, an inversely proportional relation with accuracy level was observed where lower levels of accuracy result in higher numbers of trials to criterion. The relation is somewhat weak given substantial overlap across groups. However, a strong difference was observed between the groups exposed to IF and the 100 % accuracy comparison group.

For further comparison between the groups, the proportional area under the curve (%AUC) was calculated for each participant and each phase. Data for participants in the

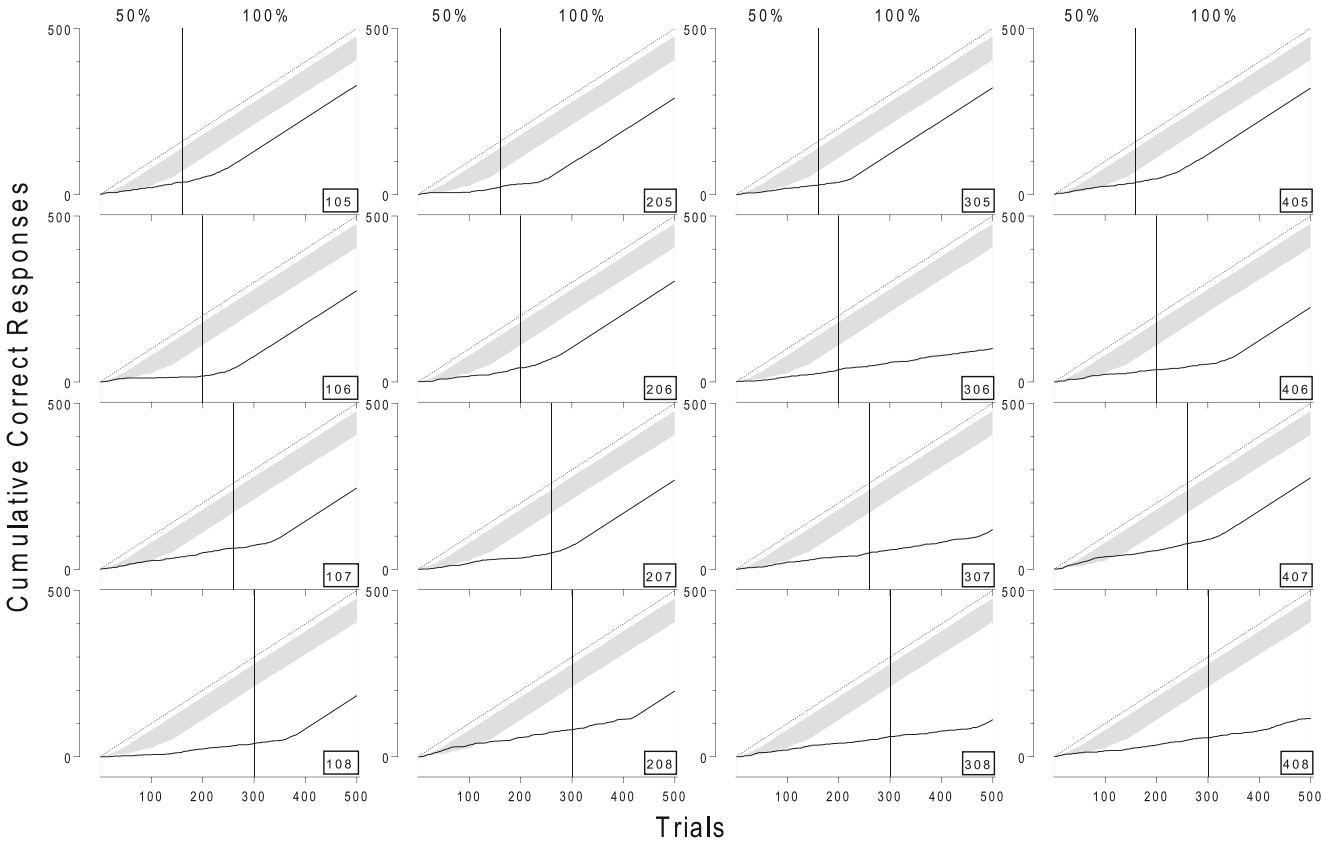


Fig. 2 Cumulative records for participants in the 50 % accuracy group

comparison group were arbitrarily split into two phases using the same method as for the experimental groups, even though these participants were never exposed to errors. That is, all participants in the 100 % accuracy condition were randomly assigned to a phase change at trial 160, 200, 260, or 300. The top panel of Fig. 5 presents the resulting data from this calculation across all participants in IF, and data from AF are presented in the bottom panel. Visual inspection of these graphs reveal a linear relation between feedback accuracy and %AUC with higher levels of accuracy associated with higher rates of acquisition in both conditions. However, substantial overlap in the data between groups was observed. These data also further depict higher levels of acquisition associated with higher levels of feedback accuracy. Although %AUC values increased for all groups in AF compared to IF, data from participants exposed to lower levels of feedback accuracy remained low, illustrating delays and occasional failure to acquire the task.

Figure 6 displays a dot plot of all of the indices of discrimination response strength for incorrect responses calculated as conditional probabilities. Generally, participants emitted incorrect responses at or below chance levels (index of 0.2), suggesting that the majority of possible incorrect responses was not strengthened as a result of exposure to IF. However, 14 data points in the 25 % group and 4 data points from the 50 % group fell above 0.5, suggesting that in the presence of many feedback errors, some incorrect discrimination responses were acquired. A pattern was obtained between indices of incorrect discrimination strength and accuracy level with fewer data points falling substantially above chance levels with higher levels of feedback accuracy.

To augment the visual inspection of these graphs and to determine whether significant differences were obtained, a nonparametric ANOVA (Kruskal-Wallis) was conducted. The results of the test showed that the differences between groups were statistically significant in both phases (IF: $H(3)=48.17, p<.0001$; AF: $H(3)=45.53, p<.0001$). A multiple comparison posttest was conducted to determine which of the groups differed significantly. During both conditions, significant differences were obtained between all groups except between 25 % and 50 %, and 75 % and 100 %. These data suggest that the number of errors to which a participant is exposed differentially influences learning when comparing relatively many errors to relatively few errors. The lack of differentiation between few errors and no errors suggests a ceiling effect in which some less-than-perfect level of feedback accuracy does not result in statistically significant differences in acquisition compared to perfectly accurate feedback. For example, for some participants, exposure to 75 % accuracy did not appear to hinder acquisition of the task. Given the distinct difference between participants 111 and 112, who did not acquire the task, and the other participants in the 75 %

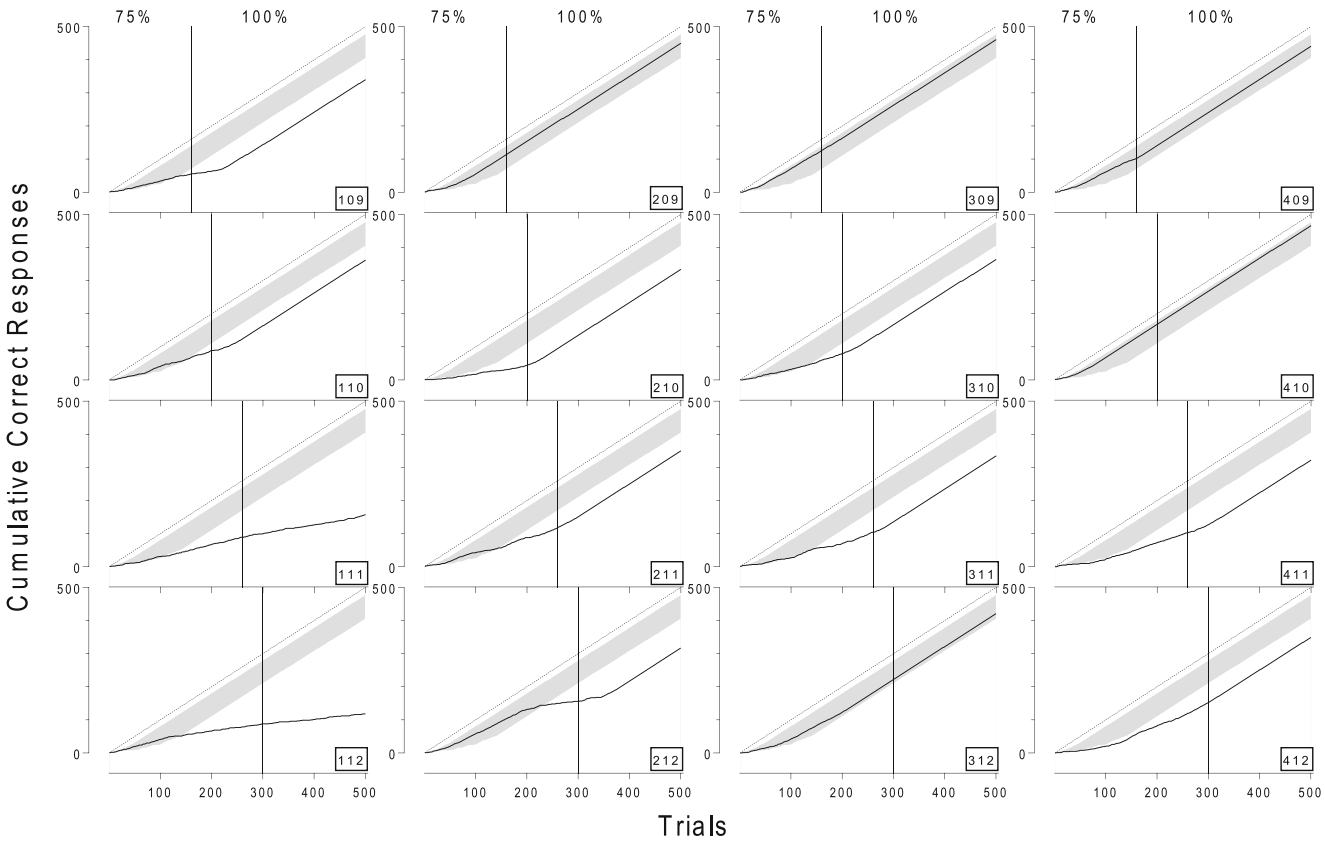


Fig. 3 Cumulative records for participants in the 75 % accuracy group

group, it may be that the presence of even a few errors can significantly impact learning for some individuals. However, it may also be that the exclusionary criterion was too conservative and did not capture some dimension of nonattending that influenced observed outcomes.

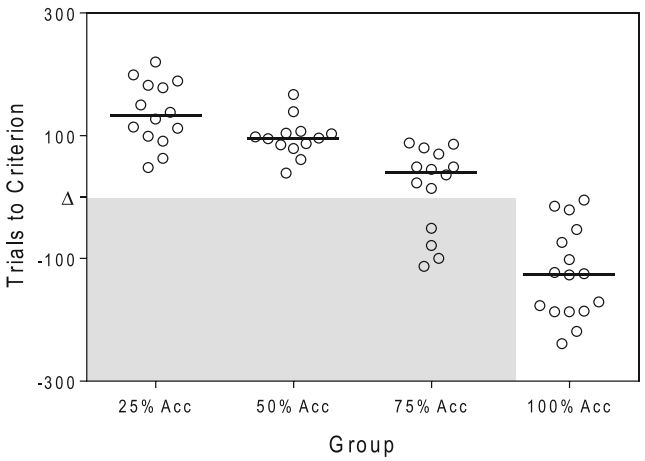


Fig. 4 Number of trials in accurate feedback before mastery criterion of 15 consecutive correct responses was met, displayed by group. Each data point represents one participant, and the horizontal solid lines are group medians. Data points below the delta on the *y*-axis, in the shaded region, represent mastery before the phase change from inaccurate feedback to accurate feedback

The findings from Study 1 are consistent with those in Hirst et al. (2013). Specifically, both studies found at least a weak linear relationship between the level of feedback accuracy and the rate of acquisition of the task. Furthermore, both studies found that groups exposed to adjacent levels of feedback accuracy showed a significant overlap in performance. Study 2 was designed to determine the generality of these findings by evaluating whether these results would also generalize to a different population, setting, and task.

Study 2

Method

Participants and Setting The participants in this study were four typically developing preschool age students (2 males, 2 females) recruited from a Montessori school located in a large Midwestern suburb. All participants were 4 or 5 years old. In addition to consent provided by parents or guardians, the institutional review board required that we provide participants an opportunity to give verbal assent. Participants gave assent on a majority of occasions but occasionally withheld it when other appealing events were taking place at the school (e.g., a waterslide was set up on the playground). Sessions

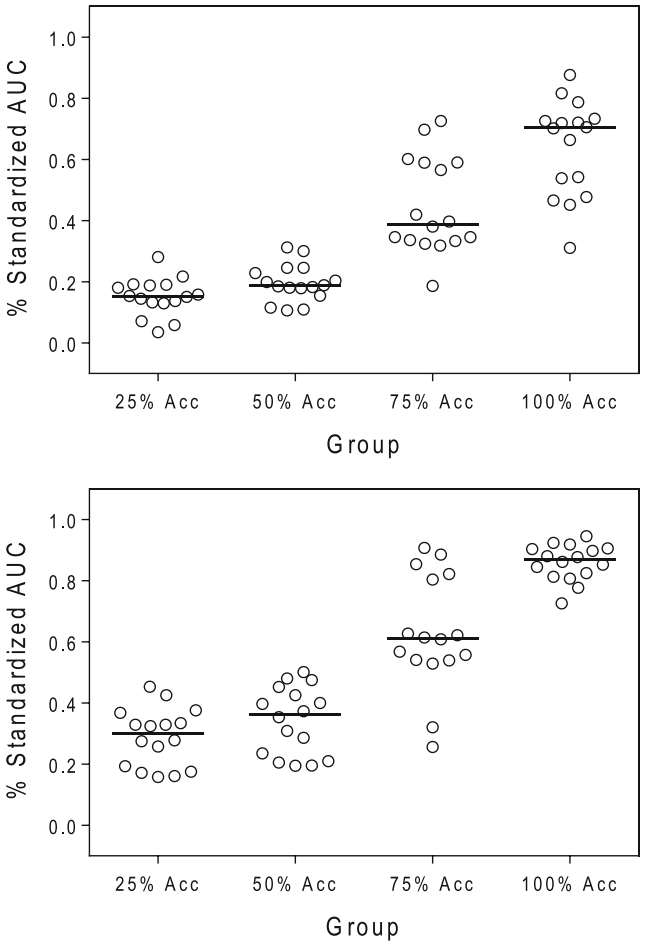


Fig. 5 Standardized percent area under the curve during the inaccurate feedback condition (top panel) and accurate feedback condition (bottom panel) by group. Solid horizontal lines are the median for each group

were conducted in a small room containing a large table and several chairs.

Materials Materials for the present study consisted of four receptive tasks. Categories of stimuli were chosen for the tasks that (a) included at least five distinct stimuli to serve as choice options, (b) prior exposure to the stimuli was unlikely, and (c) the stimuli were not part of regular curricular learning at the school. Stimuli for each task were presented on 8.5" by 11" (21.6 cm × 27.9 cm) sheets of colored paper. Each sheet was placed in a clear plastic sheet protector and mounted on a three-ring binder. The first receptive task consisted of the identification of a country in Asia on a blank outline map printed on pink paper. To simplify the task, all elements of the map were blacked out except for the shapes of five countries; one target and four distracters. A trial for this task consisted of an instruction, provided by the experimenter, to color in the target country using a dry-erase marker. The second task consisted of the identification of an aquatic invertebrate. Drawings of five insects were printed on a green sheet of paper. One served as the target, and the other four were distracters. During each trial for this task, the experimenter instructed the participants to find the target insect and color it in or circle it. Third, participants were instructed to identify, by tracing with a dry-erase marker, a river in Europe. The stimuli for this task consisted of an outline of the European continent and five thick, black lines representing major rivers. Finally, the fourth task developed for the study required participants to identify an image of a moon orbiting Jupiter. A colored picture of Jupiter was printed on a white sheet with five images of its moons oriented around it. In a trial, the experimenter instructed the participants to find the target moon and circle it.

The binders contained 24 copies of the stimuli for each task, totaling 96 pages. The order of the pages in the binder and the order of the tasks were determined by a pseudorandom sequence such that tasks were not presented twice consecutively.

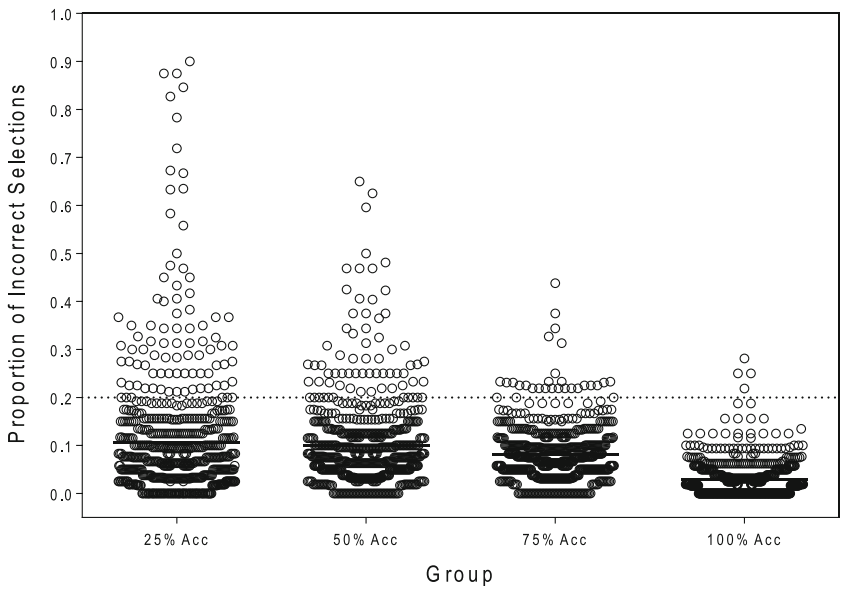


Fig. 6 Index of strength of incorrect discriminations as represented by the proportion of incorrect selections of comparison stimuli in the presence of all sample stimuli for all participants by group. The dotted horizontal line at y=0.2 is an approximation of chance levels of responding

Each task was presented an equal number of times. The order was randomized using a random number generator found at www.random.org, and two binders were created with different sequences.

Dependent Variables and Response Measurement The dependent variable for the present study was the participants' responses to the task. Correct responses were defined as identifying (i.e., coloring, tracing, circling, or pointing) the item corresponding with the instruction given at the beginning of the trial. Correct responses were graphed as a cumulative total. The mastery criterion applied to each task was 10 consecutive correct responses.

Interobserver Agreement and Procedural Fidelity Interobserver agreement and procedural fidelity measures were collected for at least 30 % of trials for each condition and for each participant. Interobserver agreement was collected by an independent observer, and agreement was calculated for participant responses as the number of agreements—defined as both observers recording the same response—divided by the number of agreements plus disagreements, represented by a percentage. Interobserver agreement averaged 99.67 % (range: 96.7 %–100 %). To measure procedural fidelity, an independent observer recorded whether the experimenter implemented the procedure correctly, which consisted of four items per trial. The four steps included (1) presenting the correct materials, (2) presenting the correct discriminative stimulus, (3) providing feedback according to the programmed schedule, and (4) presenting a token when appropriate. Procedural fidelity was measured as the percentage of steps performed correctly by the experimenter during a session out of the total number of steps applicable to the session. Procedural fidelity averaged 99.75 % (range: 98.4 %–100 %).

Experimental Design and Procedures The effects of four levels of feedback accuracy on learning were assessed using a multielement design embedded within an ABC design. Each of the four receptive tasks was associated with one of four levels of feedback accuracy: 25 %, 50 %, 75 %, or 100 %. The task associated with each level of accuracy was counterbalanced across participants. Sessions were conducted once per day and 2 to 3 days per week. A session was defined as a brief introduction, followed by 6 to 12 blocks of four trials (one per task; 24 to 48 trials total) presented in an interspersed format, and an opportunity for participants to trade in tokens earned during the session for an item from a "store," which was populated based on the results of a preference assessment. Sessions lasted between 15 and 20 min.

Preference Assessment and Token Economy A multiple stimulus without replacement preference assessment (Deleon and Iwata 1996) was conducted individually with each participant. Ten leisure items were included in the assessment, which were selected for academic or age appropriateness. The top five preferred items for each participant were used to populate a store. During a session, participants could earn tokens in the form of small, colored beads, which could be exchanged for items from the store at the end of the session. The prices of the items in the store were set based upon the ranking of the items in the preference assessment. The highest preferred item was worth 25 tokens (maximum per session) and other item prices were set such that each item was worth five less tokens than the item ranked one spot higher during the assessment. The preference assessment was repeated during the study if the participant asked for an item not presently available in the store or if the participant did not show interest in the items in the store during a previous session.

Baseline The purpose of the baseline condition was to assess whether the participants had already acquired the responses to the tasks. Each trial in baseline consisted of four components. First, the experimenter opened the binder and presented the first task stimulus. Next, the experimenter delivered an instruction to identify the target item for that task (e.g., "Please trace the Danube River with your marker"). Third, the experimenter allowed up to 5 s for a participant response. Last, feedback and a token were delivered on a variable time schedule for participation (i.e., making any response to the task and cooperating with instructions). During baseline, the token delivery was paired with a praise statement not associated with the task (e.g., "I like how you are sitting so nicely"; "Thanks for cooperating and being a good student!"). Specific feedback was not delivered for correct or incorrect responses, and errors were not corrected. When participants inquired about the correct answer, the experimenter informed participants "I can't tell you yet. We have to see if you already know it first." Participants remained in baseline until data were stable with performance at or below chance levels, or showed a consistent pattern of correct and incorrect responses for three or more consecutive presentations of each task.

Inaccurate Feedback The purpose of the inaccurate feedback condition (IF) was to assess the effects of feedback accuracy on learning. The procedure during IF was identical to baseline except feedback was delivered following all responses to the tasks. The form of the feedback varied depending on the level of accuracy associated with the task. During trials for the task associated with 100 % accuracy, all correct responses were followed by a praise statement and one token. Incorrect responses were followed by neutral feedback (e.g., "Nice try.") and no tokens. During trials for the task associated with 75 % accuracy, feedback was delivered identically to the 100 % accuracy task except that on one of four trials, inaccurate feedback was given. Inaccurate feedback was defined as

following a correct response with a neutral statement and no tokens, or following an incorrect response with a praise statement and one token. During trials for the 25 % accuracy task, the procedure was reversed such that three of four responses were followed by IF. Finally, during trials for the 50 % accuracy task, the type of feedback was scheduled such that half of responses were followed by IF. Errors were not corrected in this condition regardless of accuracy level.

Accurate Feedback The purpose of this condition was to assess whether prior exposure to IF influenced learning after accuracy improved. The procedure for the accurate feedback (AF) condition was identical to the procedure for IF except that feedback errors were no longer committed. That is, all correct responses were followed by a praise statement and one token while all incorrect responses were followed by a neutral statement and no tokens.

One-Month Maintenance The purpose of the maintenance condition was to determine whether maintenance of acquisition would vary as a function of prior exposure to IF. Sessions were conducted approximately one month following the last session of AF. The procedures during this condition were identical to those of baseline.

Data Analysis Data on correct responses for each task were graphed as the cumulative total of correct responses. A slope of 1.0 represents perfect acquisition of the task while slopes lower than 1.0 represent lower rates of acquisition. An analysis was also conducted to determine whether the failure to acquire tasks during IF was associated with specific patterns of responses. The frequency of each possible response (one target, four distracters, and responses other than the five programmed comparison stimuli) was counted for each task. An index of discrimination strength was calculated similarly to Study 1 as a conditional probability (P[R|S]) for each response being emitted in the presence of each discriminative stimulus (the instruction presented by the experimenter and the task stimuli). Given that there were five programmed comparison stimuli for each task, an index of 0.20 approximates chance levels. Indices substantially above 0.20 for incorrect responses may represent the acquisition of an incorrect discrimination response.

Results and Discussion

The top panel of Fig. 7 depicts data for Participant A. During baseline, correct responses were emitted near or below chance levels, indicating the tasks had not already been mastered. During IF, the task associated with 100 % accurate feedback was mastered at Block 19, and the tasks associated with IF continued with a low proportion of correct responses. No observable differentiation between the IF conditions was observed. During AF, acquisition of the tasks previously associated with inaccurate feedback occurred. The first to meet mastery criterion was the task previously associated with 50 % accuracy after 22 trials at Block 50. The 25 % accurate feedback task was mastered after three additional trials at Block 53. Last, the task previously associated with 75 % accuracy was mastered after 26 trials at Block 54. After a period of 4 weeks, a maintenance probe was conducted. For all tasks except the task previously associated with 25 % accuracy, correct responses were emitted for all blocks. For the 25 % accuracy task, correct responses were emitted during the first three blocks, followed by three incorrect responses during the last three blocks.

Data for Participant B are depicted in the bottom panel of Fig. 7. There were no correct responses emitted during baseline, indicating that the tasks had not been previously learned. During IF, Participant B did not demonstrate acquisition of any tasks initially. Correct responses were emitted near or below chance levels for approximately 144 blocks. At Block 119, the neutral statement following an incorrect response to the 100 % accuracy tasks was modified to "No, that's not it. Maybe try picking a different one next time." This change to the procedure was made because Participant B was consistently emitting the same incorrect response and not sampling the contingencies during the 100 % accuracy task. The feedback delivered in its original format did not appear to influence her behavior. That is, receiving feedback that her response was incorrect did not result in varied responding on future trials. This pattern was not observed for the other three tasks. The block at which the feedback was changed is depicted by an asterisk on the graph. At Block 159, the mastery criterion for the 100 % accuracy task was met. No other tasks were mastered during IF. During AF, acquisition of the tasks previously associated with inaccurate feedback was observed. The 75 % accuracy task showed acquisition first, but incorrect responses were still emitted until the mastery criterion was met after 63 trials at Block 229. The task associated with 50 % accuracy was mastered after 66 trials at Block 232. The 25 % accuracy task was mastered after 68 trials at Block 234. During a maintenance probe, correct responding continued for all tasks.

The top panel of Fig. 8 portrays data for Participant C. Correct responding occurred at or below chance levels during baseline. During IF, the task associated with 100 % accuracy was mastered in 11 trials, meeting criterion at Block 19. The tasks associated with inaccurate feedback were not acquired during IF, and little differentiation between the tasks was observed. During AF, the first task to be mastered was the task previously associated with 75 % accuracy, meeting criterion in 10 trials at Block 33. The 25 % accuracy task was mastered in 19 trials at Block 42. The task previously associated with 50 % accuracy was mastered in 34 trials at Block 57. During a maintenance probe, correct responses were emitted for all tasks except the task previously associated with 50 % accuracy. No correct responses were emitted for this task.

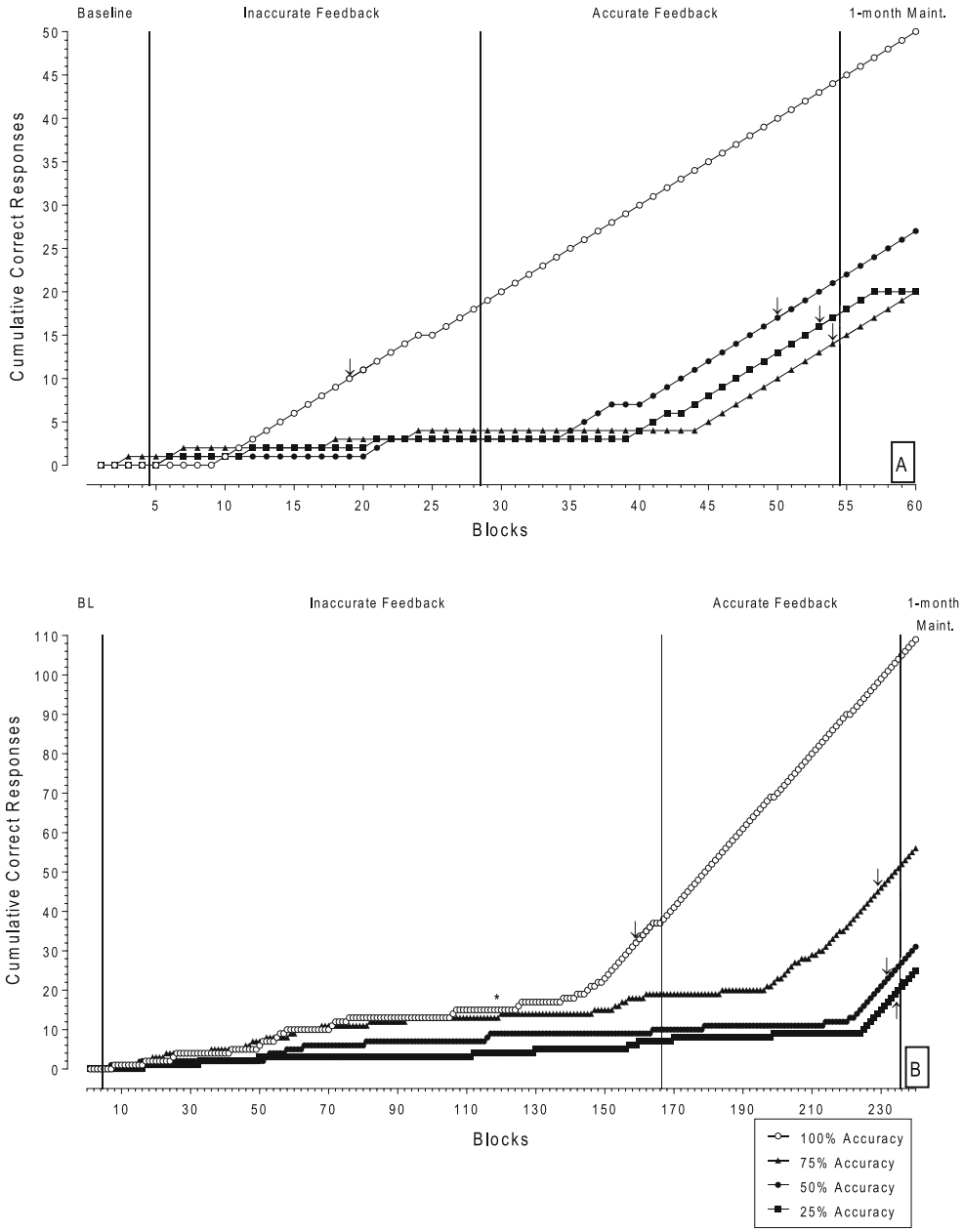


Fig. 7 Cumulative correct responses by task for Participant A (top panel) and Participant B (bottom panel). Arrows denote the block at which the mastery criterion was met for each task

Data for Participant D are displayed in the bottom panel of Fig. 8. During baseline, no correct responses were emitted. During IF, the 100 % accuracy task was mastered in 14 trials at Block 18. The tasks associated with inaccurate feedback were not acquired, with only one correct response emitted between the three tasks. During AF, the first task to be mastered was the task previously associated with 75 % accuracy in 11 trials at Block 31. Second, the task previously associated with 25 % accuracy was mastered in 14 trials at Block 34. The 50 % accuracy task was mastered last after 23 trials at Block 43. During the maintenance probe, Participant D emitted correct responses for the task previously associated with 50 % accuracy. Three correct responses out of four blocks were emitted for the 100 % accuracy task. Out of four blocks, two correct responses were emitted for the 25 % accuracy task and one for the 75 % accuracy task.

The results demonstrated a clear difference between learning under 100 % accurate feedback and any amount of inaccurate feedback. Task acquisition occurred the most rapidly and consistently when only accurate feedback was provided. However, a clear relation was not obtained for the three levels of inaccurate feedback. During IF, acquisition was largely undifferentiated between the three imperfect feedback conditions for three of four participants. At the end of IF, Participant B showed differentiation in the expected pattern, with the most correct responses emitted to the 100 % accurate feedback task and the least to the 25 % accurate feedback task, although the difference between them was minimal. In contrast, the other participants showed no

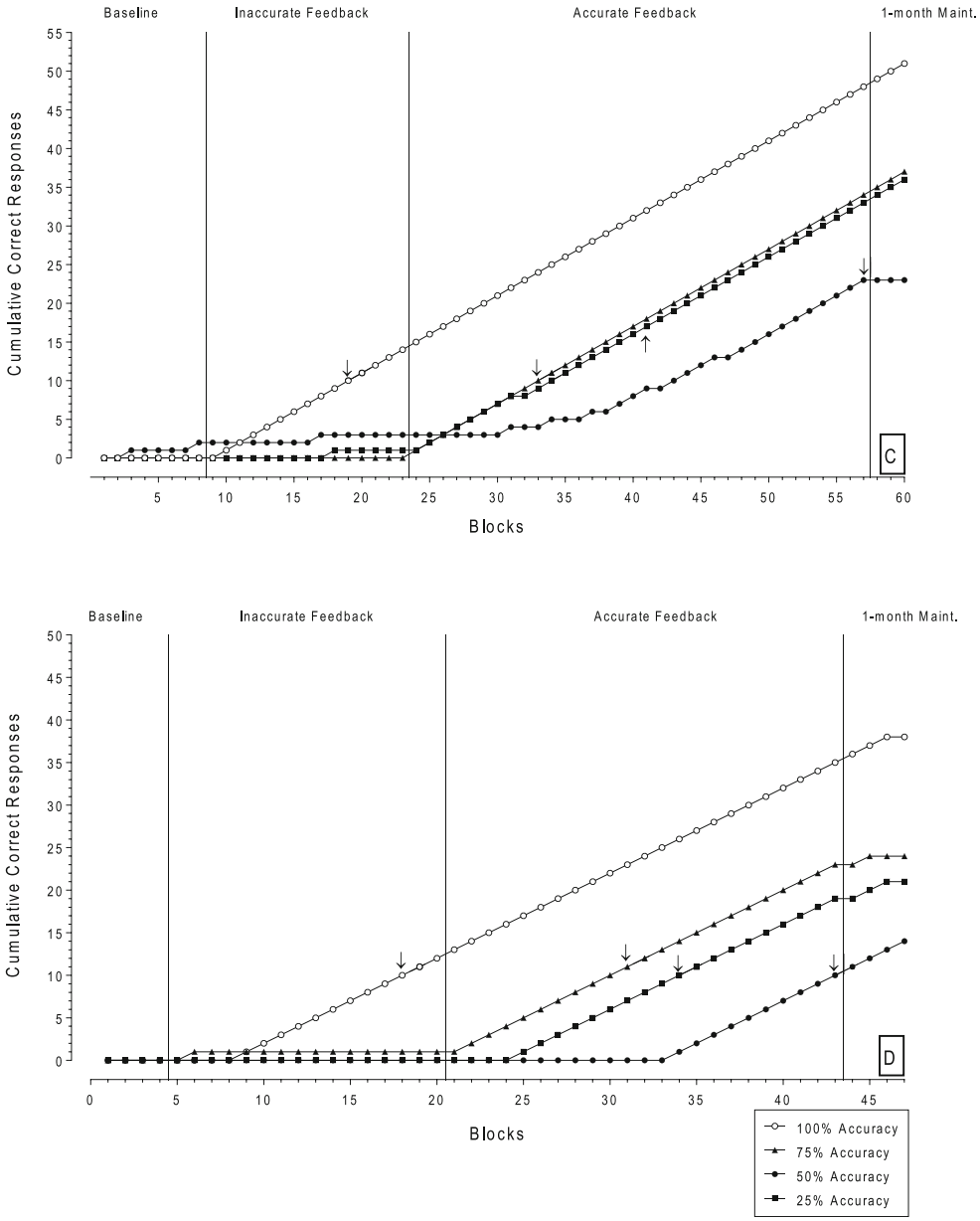


Fig. 8 Cumulative correct responses by task for Participant C (top panel) and Participant D (bottom panel). Arrows denote the block at which the mastery criterion was met for each task

differentiation or even slight differentiation, which is counter to our expectations. The relation between the level of feedback accuracy and the magnitude of delay obtained in AF before the task was mastered was also not clear. Again, Participant B acquired the three inaccurate feedback tasks in the expected order with the highest level of accuracy associated with the shortest delay to mastery (75 % accuracy). However, this pattern did not occur with the other three participants. Finally, during maintenance probes, the level of feedback accuracy previously associated with the tasks did not appear to systematically influence maintenance. Participant A failed to maintain mastery of the task previously associated with 25 % accurate feedback. However, Participant D showed the lowest maintenance of the task associated with 75 % accurate feedback and for Participant C, the 50 % accurate feedback task.

More consistent patterns were obtained on the basis of the task. When the river task was associated with any level of inaccurate feedback, maintenance of the task was low. In addition, when the river task was associated with 100 % accurate feedback (Participant B), incorrect responses were still occasionally emitted even after the task had met mastery criterion. It may be that the river task was more difficult than the other three tasks; the relevant features of the stimuli may have been less discriminable.

The results of the analysis of indices of incorrect discrimination response strength are depicted graphically in Fig. 9, with chance-level responding denoted by the reference line at 0.2. Generally, it appears that participants acquired incorrect discrimination responses to the tasks associated with inaccurate feedback. That is, participants selected the same incorrect response for which they occasionally received praise and a

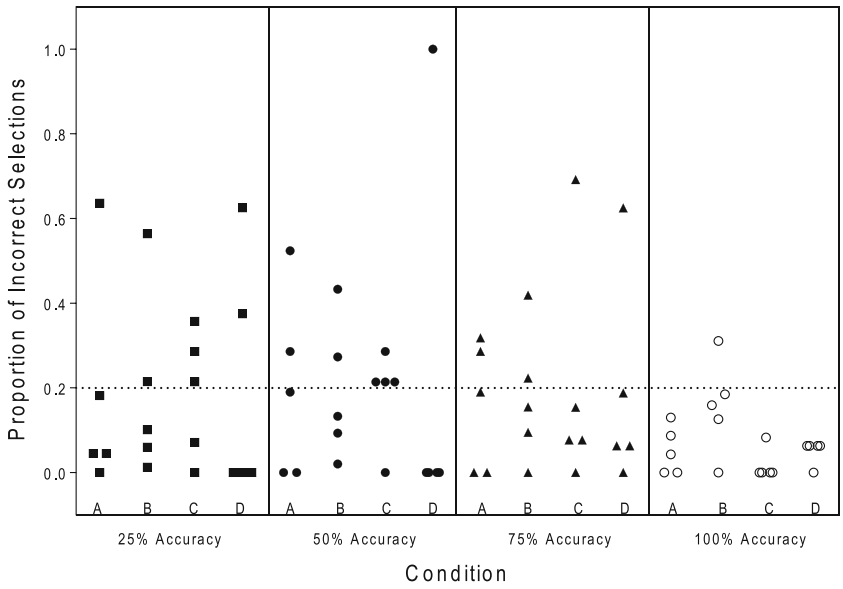


Fig. 9 Index of strength of incorrect discriminations as represented by the proportion of incorrect selections of comparison stimuli in the presence of each sample stimulus for all participants by condition. The dotted horizontal line at $y = 0.2$ is an approximation of chance levels of responding

token until AF was implemented. The analysis showed that this did not occur during the task associated with only accurate feedback.

The hypothesis described above, that learning was hindered through the strengthening of incorrect responses, was supported further by a post hoc analysis for a reinforcement effect of feedback. That is, an analysis of trial-by-trial responding revealed a hit-and-stay pattern of responding. When participants received feedback that the response they emitted was correct, they were likely to continue to select the same response on the next presentation of the discriminative stimuli. The trial-by-trial analysis is depicted visually in Figs. 10, 11, 12 and 13. Open circles represent the experimenter providing feedback that the response emitted was correct, regardless of whether it actually was correct. Closed circles represent the experimenter providing feedback that the response emitted was incorrect. Two distinct patterns of responding were obtained: (1) some participants perseverated on a response despite repeated deliveries of feedback that the response was incorrect, and (2) some participants selected a different response following only one delivery of this feedback. Although clear conclusions cannot be drawn from this analysis, the results support two possible functions of the feedback. A reinforcement function of feedback is supported by data for participants who persisted in selecting responses that were previously consequated with feedback that the response was correct. However, other participants rapidly switched responses after failing to contact positive feedback only once, suggesting that the feedback may have served a function of rule governance. The rule appeared to be: If positive feedback is delivered, then select the same response; if positive feedback is not delivered, then switch responses.

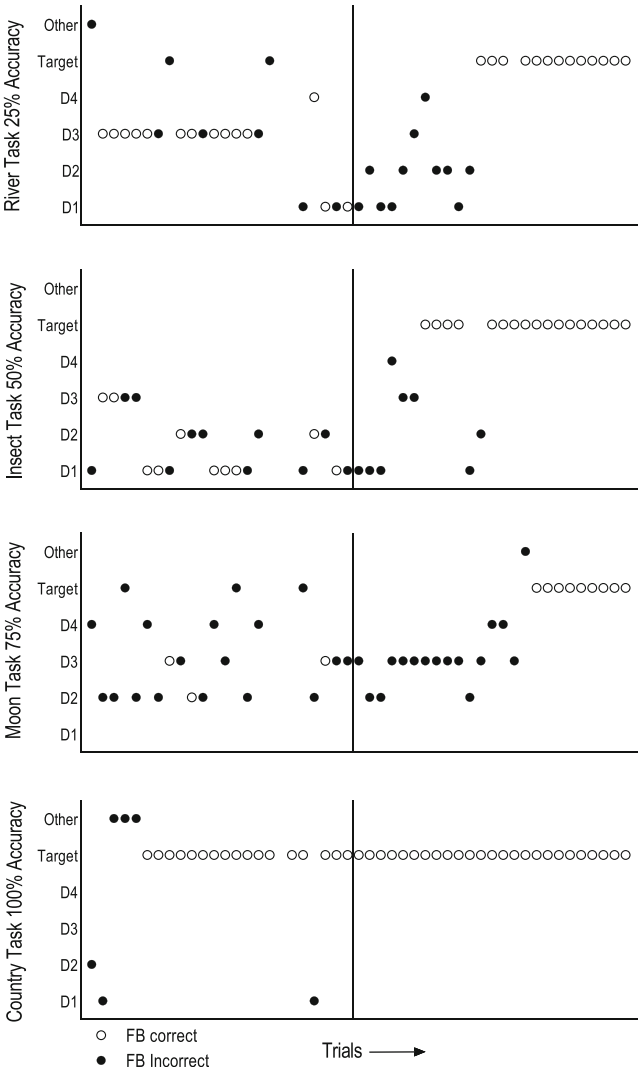


Fig. 10 Trial-by-trial analysis for Participant A, denoting selection and feedback delivered

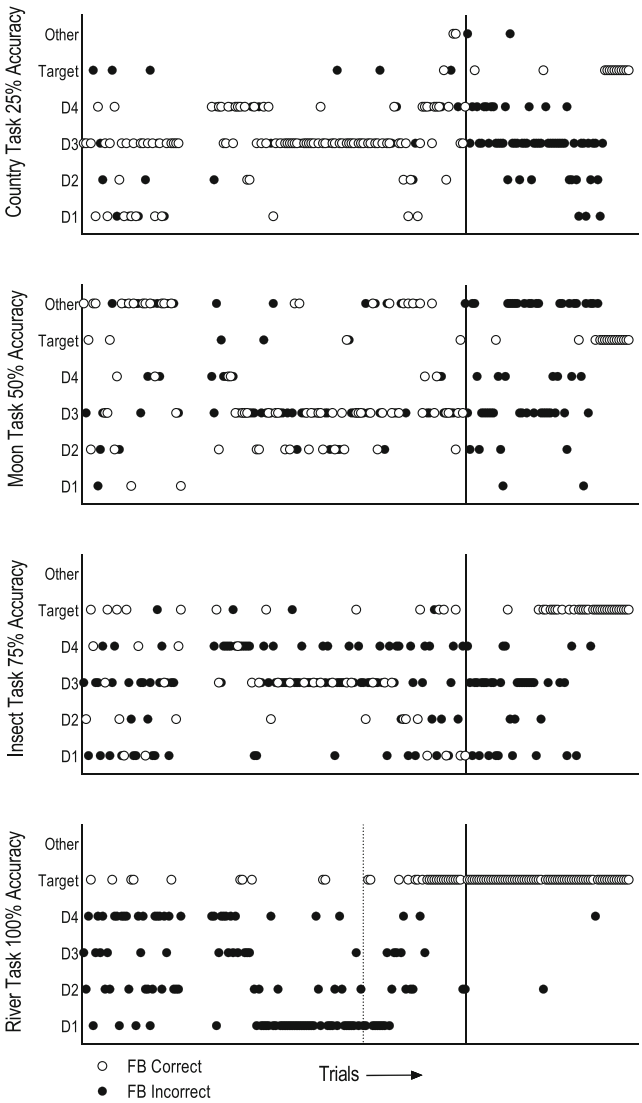


Fig. 11 Trial-by-trial analysis for Participant B, denoting selection and feedback delivered

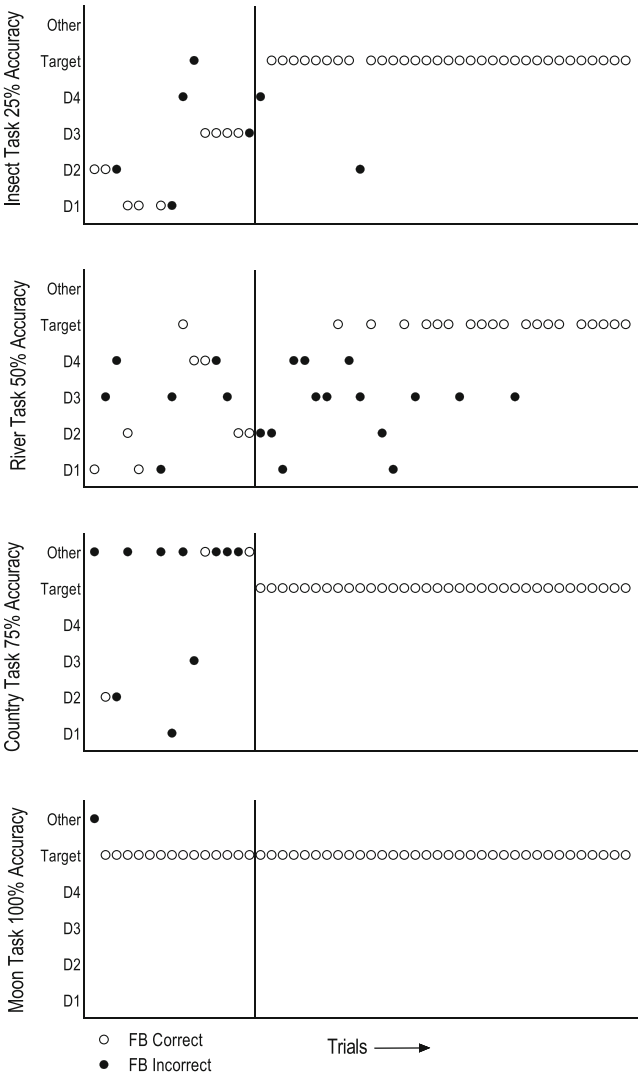


Fig. 12 Trial-by-trial analysis for Participant C, denoting selection and feedback delivered

The results of the present study are consistent with previous research and with the results of Study 1 in that the best learning outcomes were obtained only when accurate feedback was delivered. The lack of differentiation during IF between the three levels of inaccurate feedback appears to be consistent with the overlap between groups in Study 1. Although the results of the present study extend findings on feedback accuracy, a few limitations should be noted. The present study was designed to extend the findings of past research and Study 1 to a new population. However, the generality of these findings to applied settings may still be limited given the analogue nature of the study. Additionally, despite counterbalancing the assignment of accuracy level to tasks, we could not necessarily rule out the influence of carryover effects inherent to the multielement design. It is possible that exposure to IF during one task may have influenced responding to other tasks. For example, if participants developed a rule that the feedback provided by the experimenter is inconsistent or inaccurate, participants may have failed to discriminate the different feedback conditions from each other.

General Discussion

The purpose of Study 1 was to examine the effects of feedback accuracy under controlled, laboratory conditions. The purpose of Study 2 was to test the generality of the findings in a less contrived setting. The collective results of the present studies suggest that the delivery of inaccurate or inconsistently accurate feedback leads to the failure to acquire and maintain a target response. The results of Study 1 demonstrated that a weak linear relation was obtained between accuracy level and acquisition; the results of Study 2 did not support such a relation. The lack of a linear relation in Study 2 may have

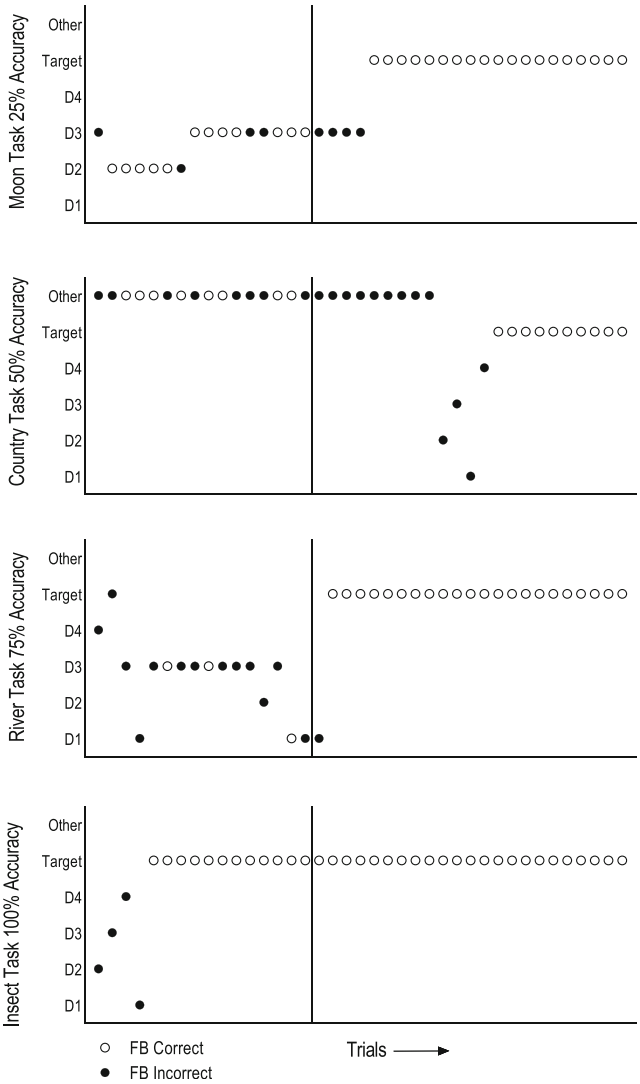


Fig. 13 Trial-by-trial analysis for Participant D, denoting selection and feedback delivered

resulted from a lack of discrimination between the tasks in the multielement design because all four levels of accuracy were assessed rapidly within teaching interactions. However, differentiation between the three inaccurate feedback conditions and the AF comparison condition was obtained.

The results consistently demonstrated that despite participants emitting occasional correct responses, and intermittently receiving positive feedback for emitting correct responses, the preceding and subsequent delivery of IF promoted variable responding over consistent adherence to a single response. The analysis of within-session data in Study 2 showed that the feedback seemed to exert strong control over responding for three of the four participants. When participants received positive feedback following a response, a strengthened tendency toward emitting the same response on the next opportunity was obtained. However, the delivery of negative feedback following a response somewhat consistently resulted in the participant emitting a different response on the next opportunity. The elastic nature of responding revealed through this analysis seems to suggest a rule-governance function for the feedback. This function of rule governance seems to be supported by participants who switched responses following one or two deliveries of negative feedback for a previously supported response. However, one participant showed a stronger persistence in responding, despite failing to receive positive feedback (Participant D). Participant B from Study 2 emitted a pattern of responding suggesting insensitivity to either form of feedback. Feedback for her may not have functioned as either a reinforcer or as a contingency-specifying stimulus.

Another conclusion drawn from the data is that the provision of accurate feedback following a period of exposure to IF did not immediately produce acquisition. That is, a persistent effect of inaccurate feedback was obtained for many participants in both studies. Nevin and colleagues (Nevin et al. 1983; Mace et al. 1990) demonstrated a persistence effect of recent reinforcement history, both in the laboratory and in an applied setting. If positive feedback served as a reinforcer, this persistence effect might account for this result. Alternatively, if the delivery of feedback served to facilitate the generation of implicit rules by participants, a carryover effect is also expected based on the relative rigidity of instructed behavior. For example, Shimoff et al. (1981) found that responding in the context of instructions was insensitive to changes in the schedule of reinforcement. However, it remains unclear whether the delay to acquisition was a direct result of the inaccurate feedback, or whether the delay was an artifact of another pattern of responding (i.e., a disruption in previous patterns of responding that failed to stabilize for some time). The molecular analysis from Study 2 lends support to both explanations. Participant A's responding following the removal of IF from the task previously associated with 75 % accurate feedback was marked with a persistence of one particular response for several trials, followed by a period of sampling other responses, and then ending by consistently emitting the correct response. However, an immediate pattern of response sampling was observed for the 25 % accuracy task for the same participant. Participant A tacted a strategy at the beginning of the penultimate session: "I'm going to try them all until I get it right." This strategy appeared to influence the duration of delay to acquisition for this participant as his systematic sampling of response options occurred spatially from left to right on the stimulus sheet, which resulted in a longer delay than would have been obtained if the order were reversed (the correct response was on the right side of the sheet). Given the higher complexity of the task in Study 1 (five sample stimuli, nine comparison stimuli), a systematic sampling of response options would likely require a substantial number of trials to complete, potentially accounting for the longer delays to acquisition obtained.

A consideration presented by the present studies is the behavioral function of the feedback provided to participants. Peterson (1982) argued that feedback is best described as a sort of "professional slang" (p. 102) and might serve any number of functions, or several, depending on history and context. Determining the function of feedback would help to understand why feedback is effective in practice. These results provide evidence for a reinforcement effect. For example, some participants in both studies who received positive feedback following responses were more likely to emit the same response again at the next opportunity. Conversely, participants who received the feedback "Incorrect" were less likely to emit the same response during the next opportunity. These changes in local response strength are in line with functional definitions of reinforcement and punishment, but the data presented here are insufficient to rule out alternative functions, such as rule governance.

These findings may also have some relevance in the area of procedural fidelity, which refers to the degree to which a trained interventionist implements a procedure (including assessment, treatment, instruction, or other protocol) as planned (DiGennaro Reed and Codding 2014; Yeaton and Sechrest 1981). Interventions commonly incorporate consequent procedures; inaccurate implementation of these contingencies may negatively influence treatment outcomes. For example, St. Peter Pipkin et al. (2010) manipulated the accuracy of reinforcement contingencies for both challenging and appropriate behaviors in the context of a behavior reduction procedure and showed that commission errors resulted in poorer outcomes, depending on a participant's fidelity history. Other interventions include a wide range of instructional practices such as discrete-trial training, precision teaching, direct instruction, academic interventions, communication training, and others. Carroll et al. (2013) systematically manipulated the implementation accuracy of a discrete trial instruction procedure, including both antecedent and reinforcement procedures, and showed that instructional errors produced delays to acquisition. The current findings offer some support for the notion that deviations in an instructional protocol that involve errors in feedback accuracy influence both immediate and later learning (history/sequence effects; see St. Peter Pipkin and Vollmer 2009), suggesting that feedback errors may constitute procedural integrity challenges.

The potential application of the present findings should be tempered given the relatively exploratory nature of the line of inquiry. Further research and modeling of applied concerns will be necessary to identify the breadth and scope of application of this characteristic of feedback. The present studies took a reverse-translational approach to evaluate characteristics of performance feedback currently not described by the extant literature. This line of inquiry would benefit from research explicitly intended to provide evidence for the hypothesized functions of performance feedback, providing a conceptual framework that would help inform future research. A logical next step in the present line of inquiry might involve determining the relative strength of control by feedback and contingencies of either reinforcement or punishment. No veritable sources of reinforcement or aversive control were programmed in the present studies. If contact with underlying contingencies either matched to the feedback, or in contradiction with the feedback, were strategically programmed, the relative influence of each could be ascertained.

Acknowledgments This investigation was supported by the University of Kansas General Research Fund allocation #2301523. The research presented in this article was completed in partial fulfillment of the requirements for the MA degree by the first author at the University of Kansas.

The authors wish to thank Derek D. Reed, David P. Jarmolowicz, and Chris Podlesnik for their invaluable feedback on an earlier version of this manuscript.

References

Alvero, A. M., Bucklin, B. R., & Austin, J. (2001). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985–1998). *Journal of Organizational Behavior Management, 21*, 3–29. doi:10.1300/J075v21n01_02.

Balcazar, F. E., Hopkins, B. L., & Suarez, Y. (1985). A critical, objective review of performance feedback. *Journal of Organizational Behavior Management, 7*, 65–89. doi:10.1300/J075v07n03_05.

Carroll, R. A., Kodak, T., & Fisher, W. W. (2013). An evaluation of programmed treatment integrity errors during discrete-trial instruction. *Journal of Applied Behavior Analysis, 46*, 379–394. doi:10.1002/jaba.49.

Chan, J. M., Lang, R., Rispoli, M., O'Reilly, M., Sigafoos, J., & Cole, H. (2009). Use of peer mediated interventions in the treatment of autism spectrum disorders: a systematic review. *Research in Autism Spectrum Disorders, 3*, 876–889. doi:10.1016/j.rasd.2009.04.003.

Cooper, M. D. (2006). Exploratory analyses of the effects of managerial support and feedback consequences on behavioral safety maintenance. *Journal of Organizational Behavior Management, 26*, 1–41. doi:10.1300/J075v26n03_01.

Daniels, A. C. (1994). *Bringing out the best in people*. New York, NY: McGraw-Hill.

Deleon, I. G., & Iwata, B. A. (1996). Evaluation of a multiple-stimulus presentation format for assessing reinforcer preferences. *Journal of Applied Behavior Analysis, 29*, 519–533.

DiGennaro Reed, F. D., & Codding, R. S. (2014). Advancements in procedural fidelity assessment and intervention: introduction to the special issue. *Journal of Behavioral Education, 23*, 1–18. doi:10.1007/s10864-013-9191-3.

Ducharme, J. M., & Feldman, M. A. (1992). Comparison of staff training strategies to promote generalized teaching skills. *Journal of Applied Behavior Analysis, 25*, 165–179.

Hirst, J. M., DiGennaro Reed, F. D., & Reed, D. D. (2013). Effects of varying feedback accuracy on task acquisition: a computerized translational study. *Journal of Behavioral Education, 22*, 1–15. doi:10.1007/s10864-012-9162-0.

Kang, K., Oah, S., & Dickinson, A. M. (2003). The relative effect of different frequencies of feedback on work performance: a

simulation. *Journal of Organizational Behavior Management, 23*, 21–53. doi:10.1300/J075v23n04_02.

Mace, F. C., & Critchfield, T. S. (2010). Translational research in behavior analysis: historical traditions and imperative for the future. *Journal of the Experimental Analysis of Behavior, 93*, 293–312. doi:10.1901/jeab.2010.93-293.

Mace, F. C., Lalli, J. S., Shea, M. C., Lalli, E. P., West, B. J., Roberts, M., & Nevin, J. A. (1990). The momentum of human behavior in a natural setting. *Journal of the Experimental Analysis of Behavior, 54*, 163–172.

Mason, M. A., & Redmon, W. K. (1992). Effects of immediate versus delayed feedback on error detection accuracy in a quality control simulation. *Journal of Organizational Behavior Management, 13*, 49–83.

Mihalic, M. T., & Ludwig, T. D. (2009). Behavioral system feedback measurement failure: sweeping quality under the rug. *Journal of Organizational Behavior Management, 29*, 55–174. doi:10.1080/01608060902874559.

Myers, W. V., McSween, T. E., Medina, R. E., Rost, K., & Alvero, A. M. (2010). The implementation and maintenance of a behavioral safety process in a petroleum refinery. *Journal of Organizational Behavior Management, 30*, 285–307. doi:10.1080/01608061.2010.499027.

Myerson, J., Green, L., & Warusawitharana, M. (2001). Area under the curve as a measure of discounting. *Journal of the Experimental Analysis of Behavior, 76*, 235–243.

Nevin, J. A., Mandell, C., & Atak, J. R. (1983). The analysis of behavioral momentum. *Journal of the Experimental Analysis of Behavior, 39*, 49–59.

Peláez, M., & Moreno, R. (1998). A taxonomy of rules and their correspondence to rule-governed behavior. *Mexican Journal of Behavior Analysis, 24*, 197–214.

Peterson, N. (1982). Feedback is not a new principle of behavior. *The Behavior Analyst, 5*, 101–102.

Prue, D. M., & Fairbank, J. A. (1981). Performance feedback in organizational behavior management: a review. *Journal of Organizational Behavior Management, 3*, 1–16. doi:10.1300/J075v03n01_01.

Rose, D. J., & Church, R. J. (1998). Learning to teach: the acquisition and maintenance of teaching skills. *Journal of Behavioral Education, 8*, 5–35.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review, 38,* 445–459.

Shimoff, E., Catania, A. C., & Matthews, B. A. (1981). Uninstructed human responding: sensitivity of low-rate performance to schedule contingencies. *Journal of the Experimental Analysis of Behavior, 36,* 207–220.

St. Peter Pipkin, C., & Vollmer, T. R. (2009). Applied implications of reinforcement history effects. *Journal of Applied Behavior Analysis, 42*, 83–103. doi:10.1901/jaba.2009.42-83.

St. Peter Pipkin, C., Vollmer, T. R., & Sloman, K. N. (2010). Effects of treatment integrity failures during differential reinforcement of alternative behavior: a translational model. *Journal of Applied Behavior Analysis, 43*, 47–70. doi:10.1901/jaba.2010.43-47.

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156–167.