Running Head: DIFFERENTIAL OUTCOMES

Quantifying the Effects of the Differential Outcomes Procedure in Humans: A Systematic Review and a Meta-Analysis

Jessica C. McCormack

Douglas Elliffe

Javier Virues-Ortega

The University of Auckland

Author Note

Correspondence concerning this manuscript should be addressed to Javier Virues-Ortega, The University of Auckland, School of Psychology, Science Centre, Building 302, Level 2, 23 Symonds Street, 1010 Auckland Central, Private Bag 92019, New Zealand. Email: j.virues-ortega@auckland.ac.nz

Abstract

We present a systematic review and a meta-analysis comparing the differential outcomes procedure to a non-differential outcomes procedure among clinical and non-clinical populations. Sixty distinct experiments were included in the systematic review, 43 of which were included in the meta-analysis. We calculated pooled effect sizes for accuracy (overall accuracy, test accuracy, transfer accuracy) and acquisition outcomes (latency, errors, and trials to mastery). The meta-analysis revealed significant medium-to-large effect sizes for all three accuracy measures (pooled effect size range, 0.57 to 1.30). We found relatively greater effect sizes among clinical populations (effect size = 1.04). The single-subject experimental literature included in the systematic review was consistent with the findings from the group studies, demonstrating improvements in accuracy and speed of learning for the majority of participants. Moderator and subgroup analyses suggest that discrimination difficulty may induce relatively larger differential effects. The results suggest that the differential outcomes procedure can be a valuable addition to reinforcement-based interventions.

Keywords: differential outcomes procedure, differential outcomes effect, meta-analysis, conditional discrimination.

Quantifying the Effects of the Differential Outcomes Procedure in Humans: A Systematic Review and a Meta-Analysis

In traditional conditional discrimination arrangements, correct performance is followed by the same reinforcers irrespective of the preceding stimulus (common-outcomes procedure). By contrast, in a differential-outcomes procedure (DOP), each discriminative stimulus is paired with a unique reinforcer. For example, selecting the green comparison after a green sample will result in a preferred edible, while selecting the red comparison after a red sample stimulus will result in access to toys (Figure 1). According to Mok, Estevez, and Overmier (2010) the DOP provides a good experimental analogue for everyday life where differential consequences are highly prevalent. For example, when driving around a commercial area screening restaurant signs, we will find that restaurant signs featuring a burger will serve burgers, those with pictures of pizza will serve pizza, and so forth. Similarly, pressing the character “E” in a keyboard results in a letter *e* on the screen, whereas pressing the character “F” will result in an *f*. In sum, the topographical dimensions of discriminative and reinforcing stimuli are often correlated in systematic ways in the natural environment.

According to the animal literature, stimulus-specific reinforcers can enhance acquisition and result in speedier learning and greater accuracy (see Urcuioli, 2005, for a review). The so-called differential-outcomes effect (DOE) was first described in animals by Trapold (1970), who trained rats in a discriminative choice task to press the left lever in the presence of a clicker and the right lever in the presence of a tone. Half the rats were trained with differential outcomes– food pellets for correct responses to the tone and sucrose for the clicker– and the other half were trained with a common outcome for both stimuli (food pellets only or sucrose only). The differential-outcomes group acquired the discrimination significantly faster than those subjects that were trained with common outcomes.

Peterson and Trapold (1982) have argued that differential reinforcers provide an additional cue that supports discrimination. They reasoned that through repeated pairings of the stimulus and the reinforcer, the stimulus comes to elicit a conditioned expectancy of the reinforcer referred to as an *expectancy-reinforcer association*. These expectancy-reinforcer associations are thought to have stimulus properties, which in turn exert control over responding. By providing a more salient or additional source of stimulus control, the expectancy-reinforcer associations facilitate discrimination learning. Trapold (1970, Experiment 2) evaluated the potential role of expectancy-reinforcer associations. Rats were pre-exposed to noncontingent stimulus-reinforcer pairings that were either consistent or inconsistent with the stimulus-reinforcer relations used in subsequent discrimination training. Rats exposed to consistent stimulus-reinforcer relations responded more accurately and learnt faster than those exposed to inconsistent stimulus-outcome relations. According to Trapold, the facilitative effect of the consistent stimulus-reinforcer presentations suggests that an existing Pavlovian conditioned response or *expectancy* may be interacting with subsequent operant discrimination processes. This antecedent influence may take the form of added discriminability of the discriminative stimulus complex (Trapold, 1970).

Alternatively, Sidman (2000) proposed that the DOE is an artifact of equivalency between stimuli that share a common reinforcer. Specifically, stimuli and responses that are reinforced with the same outcome become members of an analytic unit consisting of stimulus, comparison, and reinforcer, with the reinforcer acting as any other stimulus within the analytic unit. Sidman argued that when a pair of stimulus discriminations is taught using a common reinforcer, those stimuli may become part of a larger stimulus class joined by the common reinforcer (see also Minster, Jones, Elliffe, & Muthukumaraswamy, 2006). For example, if Stimulus A and Stimulus B are both used in the presence of Outcome C, all three may become members of the same class. Thus, the DOP could facilitate the acquisition of equivalence relations. However, on occasions when the reinforcer is not specific to the initial terms of the contingency, the reinforcer may not serve as a stimulus from which derived relations may be formed (Dube, McIlvane, McKay, & Stoddard, 1987).

In animals, the DOP has been shown to improve accuracy and speed of acquisition in conditional discriminations relative to common-outcome (Trapold, 1970) and variable-outcomes procedures (Peterson, 1984). Additionally, use of the DOP can improve retention in delayed matching-to-sample (MTS; Jones & White, 1994) and facilitate the generalization of responding to novel stimuli by way of pairing the discriminative and reinforcing stimuli in single-stimulus training (Peterson, 1984). In Experiment 1 of the Peterson study, subjects in the differential outcomes condition transferred responding to novel stimuli without any direct training, performing above 50% accuracy from the first session. Urcuioli (1990) described the DOE as “one of the most consistent and powerful effects on learning and retention of conditional discrimination” (p. 410).

The Differential Outcome Effect in Humans

Goeters, Blakely, and Poling (1992) argued for the DOP as an intervention to aid conditional discrimination learning in humans in the event that the empirical research can identify the range of conditions under which DOP is effective. A number of studies have demonstrated support for the DOE in healthy adults and typically developing children. For example, Estevez and Fuentes (2003) reported that typically developing children were more accurate in a MTS task when trained using the DOP relative to non-differential outcomes. Similarly, Miller, Waugh, and Chambers (2002) found that a DOP resulted in more accurate responding for undergraduate students learning kanji symbols than a procedure involving non-differential outcomes. However, unlike the animal literature, human studies using differential outcomes do not always report enhanced acquisition, and results may vary based on the target behavior dimension. For example, de Wit, Corlett, Aitken, Dickinson, and

Fletcher (2009) found that DOP reduced response latency but did not impact accuracy during a conditional discrimination experiment with healthy adults. Similarly, Easton (2004) and Easton, Child, and Lopez-Crespo (2011) failed to report any differences in performance during acquisition across differential- and non-differential-outcome conditions in two independent group studies. Nonetheless, participants in the DOP condition demonstrated superior performance during a transfer phase with a novel set of sample stimuli.

Conditional discrimination procedures are often used to teach new skills to individuals with intellectual and developmental disabilities (see Green, 2001, for a review). For example, conditional discrimination training is often used to establish listener responses using MTS. The DOE has been demonstrated in various clinical populations including adults and children with Down syndrome (Esteban, Plaza, Lopez-Crespo, Vivas, & Estevez, 2014; Estevez, Fuentes, Overmier, & Gonzalez, 2003a), children with autism spectrum disorder (Addison, 2006; Litt & Schreibman, 1981), adults with memory deficits (Alzheimer's Disease, Plaza, Lopez-Crespo, Antunez, Fuentes, & Estevez, 2012; Korsakoff syndrome, Hochhalter, Sweeney, Bakke, Holub, & Overmier, 2001), and adults with intellectual disabilities (Malanga & Poling, 1992) and Prader-Willi syndrome (Joseph, Overmier, & Thompson, 1997). In the clinical literature, the DOP has been shown to improve accuracy (Hochhalter et al. 2001; Malanga & Poling, 1992) and reduce the number of sessions required to reach mastery during conditional discrimination interventions (Addison, 2006; Litt & Schriebman, 1981). However, the magnitude of the effects reported in clinical studies is often inconsistent across and within participants (Chong & Carr, 2010). For example, Malanga and Poling (1992) reported more accurate performance on average in the DOP condition, but did not consistently demonstrate a DOE across repeated exposures to the DOP. Likewise, in Chong and Carr, participants did not show a consistent DOE across phases when new stimuli and responses were introduced, although they met mastery sooner under the DOP on average.

The lack of consistent findings in human studies demands a more detailed examination of the experimental conditions under which DOE is more likely to occur. A number of factors contribute to the methodological diversity in the literature: training format (MTS vs. delayed MTS), sensory modality and salience of the discriminative stimuli, modality of the reinforcing stimuli, participants' age, and training duration, among others. While these factors may vary across studies, they have not been experimentally manipulated in order to determine their impact upon the DOP.

In order to fully evaluate DOE in humans and identify areas for further analytical evaluation, we conducted a systematic review and a meta-analysis. Meta-analysis is a useful tool for aggregating the results of multiple studies and resolving uncertainty when findings vary between studies. This approach has the potential to identify complex interactions between stimuli and could help to define in quantitative terms the factors driving the diversity of effects found in the human literature, thereby providing the basis for future empirical studies. For example, meta-analytical methods can identify whether a particular characteristic or procedural element is systematically paired with a larger or smaller effect. Because it is a relatively simple procedural adaptation, the DOP has the potential to be quickly integrated as part of a range of reinforcement-based interventions as an example of *innovation through synthesis* (Mace & Critchfield, 2010).

The goal of the current study was to evaluate whether the DOP is an effective addition to discriminative training technology in applied settings. We conducted a systematic review and a meta-analysis of the currently available evidence. In addition, we were interested in identifying procedural and participant characteristics that may moderate the effect.

Method

Screening of References

In order to establish the literature base for the systematic review, we conducted a series of systematic searches unrestricted by start date and source in *PsycInfo, Medline, Cochrane Central Register of Controlled Trials, Scopus,* and *Web of Science*. The search date was June 30, 2015 (see Supporting Information for further information on the search strategy).

Systematic review inclusion criteria. We included original empirical studies meeting the following inclusion criteria: (a) the study was conducted with human participants, (b) the study used the DOP or any procedure in which different stimuli or responses are associated with different reinforcers, (c) the study involved the manipulation of a reinforcement procedure across at least two conditions or groups, (d) the study compared the DOP to a non-differential outcomes control condition or group, and (e) the study included at least one direct behavioral observation measure (e.g., percentage of correct trials). We excluded non-experimental studies, re-analyses of previously published data, and studies without a control group or control condition. We included doctoral dissertations meeting the inclusion criteria. Doctoral dissertations are known to be less vulnerable to publication bias than peer-reviewed publications (e.g., Schmucker et al., 2017).

Figure 2 summarizes the study selection process. Papers were initially screened by title and abstract in order to exclude studies that did not relate to the DOP or did not use human participants. We retrieved the remaining 71 publications for more detailed evaluation. Thirty-four publications were excluded due to lack of outcome manipulation and two publications were excluded due to duplicated data (i.e., a dissertation and published article based on the same data).

Meta-analysis inclusion criteria. In addition to the above criteria, experiments had to meet two additional criteria to be included in the meta-analysis: a minimum of five participants and fully-reported central and dispersion statistics. Six experiments were excluded for failure to meet the minimum number of participants (Addison, 2006; Chong, 2005; Experiment 3 from Maki, Overmier, Delos, & Gutmann, 1995; Malanga & Poling, 1992; Saunders & Sailor, 1979). Finally, we excluded eleven experiments that otherwise met the inclusion criteria because the dependent variables were not fully reported (deWit et al., 2009; Flores et al., 2005; Hochhalter et al., 2001; Joseph et al., 1992; Experiments 1 through 3B from Legge & Spetch, 2009; Litt & Schreibman, 1981). In summary, nine studies were excluded and two additional studies were partially excluded due to an insufficient number of participants or due to insufficient outcome reporting (Figure 2).

Data Extraction

Information from individual experiments was retrieved directly from the published work. Where required, additional information was obtained by contacting the corresponding author. If they did not respond within one month, a follow-up email was sent. We contacted 12 authors and received additional information from four. For between-group and within-subject studies we recorded all central tendency and dispersion statistics available. In addition, we extracted the following study characteristics (see Table 1 for a summary).

Research design. We classified experiments into three categories: between-groups, within-subject, and single-subject experimental design (SSED). Group experiments reported aggregated performance across individuals and experimental sessions. We classified an experiment as between-groups if participants were only exposed to one condition and comparison was made between groups of participants exposed to each condition. We classified an experiment as within-subject if the study reported means or central tendency for the whole condition and the participants were exposed to both differential and non-differential outcomes conditions. We classified experiments as SSED if the study reported the performance of individual participants using one or more single-subject experimental designs.

Participant information. Data were collected on the number of participants in each study, mean age of participants (or age range divided by two when the mean was not reported), and any diagnostic labels and medical conditions that were reported, such as intellectual disability and sleep abnormalities.

Termination criteria. We also classified experiments by performance-based or performance-independent termination criteria. Training was classified as *performance-based* when the criterion for terminating training was based on a pre-determined level of accuracy or number of correct responses, such as 100% accuracy or five successive correct responses. We classified training as *performance-independent* when the number of trials or sessions was pre-determined regardless of performance. For experiments with performance-independent criteria, we recorded the duration of training; that is, the number of trials or sessions conducted before termination (e.g., 72 trials or four sessions at 20 trials per session).

Training format. Conditional discrimination involves a specified response (often verbal) in the presence of Stimulus A and a different response in the presence of Stimulus B. Conditional discrimination training may also include a listener response. For example, participants might be shown a sample followed by an array of comparison stimuli, and must detect whether the sample is present in the array. Experiments included in the systematic review and the meta-analysis included a variety of conditional discrimination teaching formats. We classified teaching format according to two categories: matching-to-sample (MTS) and other.

In an MTS task, participants are presented with a sample stimulus and array of two or more comparison stimuli and must select the correct comparison from the array. Within the MTS tasks, we specified identity, symbolic, and delayed MTS formats. With identity MTS, the participant must select the comparison stimulus that is identical to the sample stimulus (e.g., select the green comparison when the green sample stimulus is presented). In symbolic MTS, the sample and comparison stimuli do not share physical characteristics; that is, the relationship between the sample and comparison stimuli is arbitrary (e.g., select the star when the green sample stimulus is presented). We classified the format as delayed MTS when the experimenter programmed a time delay between the removal of the sample stimulus and the presentation of the comparison stimulus.

We classified the control condition as either a *common-outcome procedure* or *variable-outcomes procedure*. In the common-outcome procedure, the experimenter arranges a single outcome for all correct responses; all stimuli are therefore followed by the same reinforcer (see Figure 1). In the variable-outcomes procedure, different reinforcers may follow correct responses on different trials, but no reinforcer is consistently correlated with a particular sample or response.

Reinforcement type. We grouped the type of reinforcement under two categories: token reinforcement and other. Experiments using token reinforcement provided participants with tokens or points for correct responding, which were later exchanged for backup reinforcers at the end of the session. All other types of reinforcement were classified under *other*. This category included tangible (e.g., toys, raffle tickets, pennies), social (e.g., praise), sensory (e.g., cheering voices), and edible stimuli (see Supporting Information for further details).

Measures of behavior. We identified six dependent variables in the relevant literature: overall accuracy, terminal accuracy, overall accuracy during a transfer phase, latency, errors, and mastery.

Three of the outcomes (i.e., overall accuracy, terminal accuracy, and overall accuracy during a transfer phase) are computed as percentage of correct trials. The anticipated effect of DOP on these outcomes is positive. Specifically, higher percentage values denote better performance for the group or condition exposed to DOP. For overall accuracy, the percentage of correct trials is calculated by the number of correct responses divided by the total number of trials. In contrast, terminal accuracy is calculated on the basis of the terminal set of trials or sessions within the training phase, or during post-training testing. Some experiments reported overall accuracy during a transfer phase after initial training. A transfer phase involved the introduction of a novel set of samples or comparison stimuli. A successful transfer has occurred when participants respond to the novel stimuli as though they were the initial stimuli. For example, in a study where the selection of a red card in the presence of an apple is reinforced, during a transfer phase, the red card would be substituted for a novel stimulus (e.g., the letter A), so that the selection of the novel stimulus in the presence of an apple is now reinforced. Transfer phases may also test for the emergence of transitivity and equivalence relations (see Sidman & Tailby, 1982).

The three remaining dependent variables have an inverse relation with the DOE. Thus, their values should be lower for the group or condition exposed to DOP. Latency is the average time taken to respond during correct trials. Latency is measured as the time elapsed since the presentation of the stimulus until the occurrence of the participant’s response. Errors are computed as either the mean number of errors made on average or the percentage of errors over a training phase or session. Finally, mastery refers to the number of training sessions or trials required to meet mastery criteria. Mastery was generally measured in trials to mastery for between-group or within-subject experiments, and sessions to mastery for single-subject experimental design.

Assessment of Methodological Quality

We used an adaption of the methodological quality standards proposed by Kratochwill et al. (2010, pp. 14-17). Adaptations were prompted by the lack of any single methodological quality rating scale that was suitable for both single-subject experimental designs and human-operant within-subject and between-subject studies. The modifications were based on the critical items of the scale by Downs and Black (1994), which has been used in previous meta-analyses of behavioral literature (e.g., Virues-Ortega, 2010).

The standards assess the quality of a study in three main areas: manipulation of independent variable; systematic measurement of dependent variable; and demonstration of effect. These criteria result in a grade for each study as meeting evidence standards, meeting evidence standards with reservations, or not meeting evidence standards. Because these standards were developed to assess methodological quality of SSED studies, we adapted or added some additional criteria in order to account for group experiments: (a) automatic data collection and calibration as an alternative to inter-observer agreement (IOA), (b) sample size brackets, (c) counterbalancing of stimuli, and (d) within-subject or a between-groups design with random allocation (see Supporting Information). The standards were applied to all experiments meeting the inclusion criteria.

Of the 60 experiments in our analysis, 40 met the evidence standard (see Supporting Information, Appendix C). Three experiments met the standards with reservations and 17 experiments did not meet the evidence standards. All of the experiments meeting the standards with reservations and sixteen of those not meeting standards failed to demonstrate systematic measurement of the dependent variables by either IOA or calibration checks. Two additional experiments did not have a sufficient number of participants to demonstrate an effect according to our sample size standard for group experiments. Finally, one study did not counterbalance the stimuli. Of those experiments that did not meet the evidence standard, eight were not included in the meta-analysis. The methodological quality scores could range from 0 to 6 for SSED, and from 0 to 13 for group experiments. The SSED had a mean methodological quality score of 5.0 (range, 2 to 6). The mean for group experiments was 9.4

(range, 4 to 13). The moderator analysis showed a positive accuracy and terminal accuracy bias in studies with relatively low methodological quality. Specifically, each point in the quality scale was associated with a 0.20 decrement in the differential outcome effect size estimate for overall accuracy and a 0.39 decrement in the differential outcome effect size estimate for terminal accuracy (accuracy: moderator analysis coefficient and standard error, -0.20 ± 0.09, 95% confidence interval, -0.39 to -0.01, *p* = 0.042; terminal accuracy: moderator analysis coefficient and standard error, -0.39 ± 0.17, 95% confidence interval, -0.76 to -0.02, *p* = 0.042). Moderator analyses by quality produced no significant findings for other outcomes (Table 2).

Inter-Rater Agreement

We conducted inter-rater agreement analyses on key aspects of the study methods including screening of references, data extraction, ascertainment of study characteristics, and assessment of methodological quality.

A secondary rater with postgraduate training in applied behavior analysis screened 25% of the original pool of references. The rater was trained according to a behavioral skills training model (Parsons, Rollyson, & Reid, 2012) comprised of the following steps: (a) the secondary rater read a short passage about the DOE and an explanation of each of the exclusion criteria; (b) the first author and the secondary rater evaluated four experiments together; and (c) the secondary rater evaluated five experiments independently (any disagreements were discussed immediately). We repeated these steps until the rater evaluated four experiments successively with perfect agreement.

Inter-rater agreement was calculated by dividing the number of agreements by the total number of experiments rated. We defined agreement as both rater and researchers treating the paper as either included or excluded. If the raters disagreed, the first author re-evaluated the paper and took this evaluation as the true value. Disagreements tended to be due to a stricter application of the criteria by the first author, particularly where there was ambiguity. Inter-rater agreement for the reference screening process was 92%.

For data extraction, the rater read a written explanation of data extracted and how they would be classified, and was trained by the first author to apply the classifications and to fill out a coding form. The first author filled out a form and described each step of the process based on a hypothetical experiment. Then, the rater and first author coded one of the experiments together and reviewed any discrepancies. The rater then coded five experiments independently, which the first author checked for agreement. Once the rater achieved 80% or greater accuracy (with errors occurring no more than once in each category), the rater continued data extraction independently for 33% of the included experiments.

For discrete categories, items were marked as in agreement if both raters coded the item identically. For continuous data, items were marked as in agreement if both raters reported the same value. In the event of a disagreement, the first author re-checked the original article and made a decision on the true value. Inter-rater agreement was calculated by dividing the total number of items in agreement by the total number of items rated. Inter-rater agreement was 93% (range, 85 to 100%).

Finally, a secondary rater evaluated the methodological quality of 36% of the experiments included in the systematic review and meta-analysis. Each of the seven criteria included in the quality assessment schedule could be marked as not applicable, meeting, meeting with reservations, and not meeting the criterion. The raters were in agreement if both gave a given category the same marking (e.g., both parties rated a study as meeting evidence standards on the dependent variable). We obtained an average inter-rater agreement of 97% (range, 71 to 100%).

Effect Calculation and Analysis

We calculated effect sizes as the standardized mean difference (*d*) between the differential outcomes condition and the non-differential control condition (or group) for between-group and within-subject experiments:

$$d = \frac{X_2 - X_1}{S_{pooled}},$$

where X_1 and X_2 are the mean for the control and DOP conditions, respectively. The pooled standard deviation was calculated as follows:

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}},$$

where *n* and *S* represent the number of participants and standard deviation in the baseline (n_1, S_1) and DOP (n_2, S_2) conditions, respectively.

The variance V_d was computed as follows:

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)},$$

where *n* represents the number of participants in the study and *d* represents the standardized mean difference (according to Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 52). We computed effect sizes for experiments with five or more participants in each condition. We used Hedges (1981) small-sample size correction by multiplying standardized mean difference and variance by *J*:

$$J = 1 - \frac{3}{4df - 1}$$

Owing to the range of study designs, populations, and training formats in the target literature, we conducted a random-effects meta-analysis (DerSimonian & Laird, 1986; Sterne, Harris, Harbord, & Steichen, 2009). The random effects model assumes that the theoretical *true effect* of an intervention may vary from study to study due to differences in procedure or population (Borenstein, Hedges, Higgins, & Rothstein, 2009). A fixed-effect model was used on occasions when the tau-squared parameter was negative (Borenstein, Hedges, & Rothstein, 2007).

In order to document the potential for publication and small-study bias, we computed Egger's regression asymmetry test (Egger, Smith, Schneider, & Minder, 1997). The Egger test evaluates the extent to which the distribution of effect sizes across experiments fits a theoretical distribution (i.e., greater variation in experiments with a smaller *n* relative to those with a larger *n*). Experiments in a meta-analysis producing a significant Egger test ($p < .05$) may be subject to publication or small-study bias.

We also evaluated the heterogeneity associated with every meta-analysis. Heterogeneity is a consequence of the clinical and methodological diversity of the experiments included in the meta-analysis (e.g., studies including different target populations or procedures) that results in a level of effect variability larger than would otherwise be expected (Sterne et al., 2009). The lower the level heterogeneity not explained by variation across studies, the more credible the results of a meta-analysis are. We used the I^2 heterogeneity index by Higgins and Thompson (2002): values between 0 and 40% indicate a low level of unexplained heterogeneity, that is, a level that should not be of concern to the validity of results, whereas values between 75 and 100% indicate considerable heterogeneity (Higgins & Green, 2011).

We conducted separate meta-analyses in defined subgroups of experiments. Specifically, we conducted subgroup analyses restricting the target literature to experiments with typically developing children and healthy adults. This approach helped to evaluate whether any effects were skewed by the inclusion of clinical populations. Subgroup analysis was also conducted on experiments involving clinical populations and, where possible, non-clinical controls from the same experiments (e.g., healthy adults or typically developing children). In addition, we conducted a meta-regression or moderator analyses using participant age, population, training format, reinforcer type, and study duration as moderators on occasions when specific moderator-outcome combinations were present in five or more experiments (Harbord & Higgins, 2008). The moderator analysis coefficients informed the quantitative relation between the predictor and the outcome. For example, a coefficient of one of a categorical predictor such as child vs. adult for overall accuracy would indicate that adults, relative to children, tend to have an overall accuracy that is higher by one effect size unit.

We did not conduct a meta-analysis of SSEDs due to insufficient participants and insufficient level of detail in the datasets available (e.g., lack of session-by-session data). Nonetheless, this literature is still discussed as part of our systematic review.

Results

The search strategy returned 442 distinct references. A total of 35 publications reporting 60 distinct experiments with a pooled *n* of 1,993 participants met the inclusion criteria for the systematic review. Twenty-seven publications used between-group (22 experiments) and within-subject (27 experiments) group designs (*n* = 1,952) (Supporting Information, Appendix D), whereas seven publications (eight experiments) used single-subject experimental designs (*n* = 41; Supporting Information, Appendix E). One additional publication reported one between-group experiment, one within-subject experiment, and one SSED (Maki et al., 1995). Forty-two experiments reported overall accuracy, making it the most commonly reported dependent variable, followed by terminal accuracy (20), latency (18), and transfer (9).

Twenty-six of the experiments reviewed evaluated DOP in children, 34 in adults, and three in older adults. Most of the experiments included in our analysis were conducted with typically developing children and healthy adults (44 out of 60 experiments), with 16 experiments conducted with typically developing children (mean age 6 years, range, 4 to 11 years) and 28 with healthy adults (mean age 23 years, range, 18 to 76 years). Sixteen experiments evaluated children and adults with various clinical conditions (mean age 26 years; range, 3 to 86 years). Nine of the experiments evaluated adults with intellectual disabilities (Down Syndrome, Prader-Willi Syndrome, and mental retardation) and children with autism spectrum disorder and intellectual disabilities. Four experiments recruited adults with memory impairments of various causes (Korsakoff syndrome, Phase 4 Alzheimer's disease, age-related memory deficits, and sleep deprivation). Finally, three experiments involved profoundly deaf children and children born prematurely.

Fifty of the experiments used performance-independent termination criteria, with a median length of training of 72 trials (see Table 1). Most of the experiments used either symbolic or identity MTS (35), with delays used in 39 of the experiments. Tokens were used in 15 experiments, with the remaining experiments using other reinforcers (e.g., praise, edibles).

The differential outcomes effect was evaluated with clinical populations in 16 experiments, including between group and within subject group experiments (8) and single-subject experiments (8). The total pool of clinical participants was 182, with 33 participants (see Supporting Information) from the SSED literature. Nine of these experiments were conducted with children and eight with adults, and one study was conducted with both adults and children (Estevez et al., 2003a). Clinical experiments reported overall accuracy (11), terminal accuracy (2), transfer (2), latency (6), and sessions to mastery (5). Only three experiments reported session-by-session data that would have been amenable to meta-analysis (Chong, 2004, Experiments 1 and 2; Addison, 2006).

Tangible or non-tangible immediate reinforcement was used in 14 of the experiments, while token reinforcement was used in four experiments. Identity-MTS was used in four of the group designs with symbolic-MTS and conditional discrimination each used in two experiments. Six of the experiments used delayed-MTS. The most common format used in the SSED was conditional discrimination (6), followed by symbolic MTS (2). Delays between the sample and comparison were only used in one study (Hochhalter et al., 2001), in which they were used in combination with identity-MTS.

Forty-three between-group and within-subject experiments met the additional inclusion criteria to be included in the meta-analysis. Our analysis revealed a significant main effect of the DOP for overall accuracy (*ES* = 0.57, 95% confidence interval [CI] 0.26 to 0.66, $p < 0.001$), terminal accuracy (*ES* = 0.80, 95% CI 0.22 to 1.37, $p = 0.006$), and transfer (*ES* = 1.30, 95% CI 0.53 to 2.08, $p = 0.001$). By contrast, we could not establish a significant effect for latency (*ES* = -0.32, 95% CI -0.76 to 0.12, $p = 0.15$*)* and errors (*ES* = -0.09, 95% CI: -1.07-0.89, $p = .86$). The i^2 values were lower than 10% for all five variables, indicating a low level of unexplained heterogeneity, with the greatest heterogeneity found for transfer (I^2 = 6.5%). Egger's test provided no evidence of publication or small-study bias in any of the outcomes evaluated ($p > .1$).

The moderator analysis revealed a significant interaction between overall accuracy and reinforcer type (moderator analysis coefficient = 0.42 95% CI 0.04 to 0.80, $p = 0.032$) (Table 2). Specifically, the effect of DOP was larger in experiments using tokens (*ES* = 1.17; 95% CI 0.51 to 1.83, $p < 0.001$) than in experiments using other forms of reinforcement (*ES* = 0.40; 95% CI 0.05 to 0.75, $p = 0.024$). In addition, we identified a number of trends that did not reach statistical significance: latency and population (moderator analysis coefficient = -2.07, 95% CI -4.28 to 0.14, $p = 0.065$), transfer and termination criteria (moderator analysis coefficient = *1*.07, 95% CI: 0.01-0.12, $p = 0.11$), and overall accuracy and teaching format (moderator analysis coefficient = *1*.29 95% CI: -0.07-0.64, $p = 0.12$).

The moderator analyses revealed significant effects involving format and termination criteria. The effect of DOP on overall accuracy was greater in experiments using delayed

MTS relative to experiments using MTS without a delay (delayed MTS, *ES* = 0.59, 95% CI 0.26 to 0.92, *p* < 0.001; no delay: *ES* = 0.42, 95% CI -0.47 to 1.30, *p* = 0.36). Similar results were found for terminal accuracy when delayed MTS was compared to other conditional discrimination procedures (delayed MTS, *ES* = 0.85, 95% CI 0.22 to 1.48, *p* = 0.009; no-delay procedures, *ES* = 0.60, 95%CI -0.80 to 2.00, *p* = 0.41). The pooled effect size for overall accuracy was also greater among the experiments using symbolic MTS (*ES* = 0.81, 95% CI 0.36 to 1.26, *p* < 0.001), whereas the experiments using identity MTS (*ES* = 0.49, 95% CI -0.20 to 1.18, *p* = 0.16) and conditional discrimination (*ES* = 0.29, 95% CI -0.24 to 0.82, *p* = 0.28) found a relatively lower overall accuracy.

Differential Outcome Effect in Children

We conducted a subgroup analysis restricting the meta-analysis to typically developing children below 18 years of age (Figure 3). The 17 experiments reported overall accuracy (14), terminal accuracy (3), transfer (6), and latency (1) as outcomes of the DOP. Ten of the experiments used token reinforcement in which children were provided with tokens specific to the sample stimuli for correct responses, which were then exchanged for either toys or food items at the end of the session. All 17 experiments used delayed-MTS with symbolic MTS (15) or identity MTS (2).

The results indicated that the effect of DOP among the child population was significant for overall accuracy (*ES* = 0.69, 95% CI: 0.15 to 1.23, *p* = .01) and transfer accuracy (*ES* = 1.39, 95% CI: 0.55 to 2.23, *p* = .001). The effect was not significant for terminal accuracy (*p*>.05).

Differential Outcomes Effect in Adults

All of the experiments involving healthy adults used non-tangible immediate reinforcers, such as sensory rewards (music or picture scenes) and textual praise. Symbolic MTS was used in eight experiments and identity MTS in one experiment. Thirteen of the experiments used non-MTS conditional discrimination and presence/absence responding. Delayed-MTS was used in 14 out of 23 experiments. Three experiments also used distractor tasks (such as articulatory suppression) in combination with delays to increase task difficulty.

We conducted subgroup analyses restricting the meta-analysis to healthy adults (Figure 4). The 23 experiments conducted with healthy adults reported overall accuracy (14), terminal accuracy (7), latency (15), and errors (3). Transfer was not reported in any of the experiments involving adult population. The DOP had a favorable effect over terminal accuracy (*ES* = 0.81, 95% CI: 0.05 to 1.57, *p* = 0.036), whereas no significant effect was established for overall accuracy, latency, and errors (*p* > .1).

Differential Outcomes Effects in Clinical Populations

Seven experiments were included in our analysis of the DOE in clinical populations (see Figure 5). The differential outcomes procedure had a large significant effect upon performance accuracy (i.e., percentage of correct responses) for clinical populations (*ES* = 1.04, 95% CI: 0.36 to 1.71, *p* = 0.003). For comparison, we also calculated the pooled effect size from non-clinical controls in six of the seven experiments (Esteban et al., 2014b; Lopez-Crespo, Plaza, Fuentes, & Estevez, 2009; Martella, Plaza, Estevez, Castillo, & Fuentes, 2012; Martinez, et al., 2012; Plaza et al., 2012). The control groups were exposed to the same procedures and conditions as the clinical participants in the same experiment. The pooled effect for the control groups was negligible (*ES* = 0.03, 95% CI: -0.72 to 0.80, *p* > .1).

Single-Subject Experiments

Our systematic review suggests that the differential outcome effect established in the meta-analysis was consistent with the findings in single-subject experiments targeting clinical populations. Specifically, 17 out of 27 individuals evaluated performed better in the differential outcomes condition, five out of the 27 participants performed better in the non-differential outcomes condition, and five showed equal performance across both conditions (see Supporting Information, Appendix F). The differential outcomes procedure was successful in producing more efficient and accurate acquisition in nearly two thirds of the participants in the SSED literature.

Discussion

We conducted a systematic review and a meta-analysis with the aim of evaluating the pooled effect of the DOP in humans and the procedural variations that are correlated with stronger effects. We found that the DOP positively affects the performance of various human populations in conditional discrimination experiments as measured by improvements to terminal and overall accuracy, and by improvements in generalization to novel stimuli or transfer. The DOP had no impact on other measures of learning efficiency, such as response latency or percentage of errors in group-based experiments, and there were insufficient data available to analyze a possible DOE on sessions or trials to mastery as dependent variables.

The moderator analysis revealed greater overall accuracy among experiments using token reinforcement relative to those using other reinforcers. Studies incorporating delayed and symbolic MTS had numerically higher overall accuracy with *p* values just below the threshold for statistical significance of a moderating effect. Studies with relatively lower methodological quality standards tended to over-report overall and terminal accuracy effect sizes. No other moderating effects were demonstrated.

In order to evaluate the generalizability of our findings across several target groups, we conducted subgroup analyses by limiting the pool of studies to typically developing children, healthy adults, and clinical populations.

The Differential Outcomes Effect among Typical Children and Adults

Separate analysis by age group demonstrated the DOE among typically developing children (overall accuracy, transfer) and healthy adults (terminal accuracy). However, the effect of the DOP on the reported outcomes was not consistent across the two age groups.

Specifically, adult experiments showed a favorable effect on terminal accuracy, but not for overall accuracy and transfer. For typically developing children, DOP had a favorable effect on overall accuracy and transfer, but not terminal accuracy. Thus, our analysis suggests that the DOE may be apparent sooner among children but only at the end of training (terminal accuracy) among adults. These results may be a byproduct of the inconsistent distribution of dependent variables across age groups (e.g., few studies with children reported terminal accuracy while few adult studies reported transfer). Nevertheless, the asymmetric distribution of outcomes across age groups was not detrimental to the DOE on overall accuracy, which was widely reported in experiments with participants of all ages. Additionally, these differences were not reflected in our moderator analysis, although the results may have been skewed by the inclusion of adult clinical populations, which demonstrated larger effects.

Procedural differences between studies involving adults and children may account for the differences found between those populations. A number of studies with children involved token reinforcement, which the moderator analysis identified as a DOE moderator. Our results are compatible with several interpretations: it may be that token reinforcement is more effective because it was implemented with younger participants, or that younger participants performed better because of token reinforcement. In addition, age may be correlated with another unknown factor to which increased performance can be attributed. For example, children may be less accurate when acquiring response chains than older adults (Baldwin, Chelonis, Prunty, & Paule, 2012). As children take longer to demonstrate highly accurate performance, differences between DOP and non-differential procedures (NDP) may be more discernible during acquisition in child populations. Conversely, the lack of effect for adults in acquisition may be due to a ‘ceiling effect’: because adults perform more accurately there is less room for improvement and therefore differences would be less likely to be observed during acquisition. Experiment 1 in Maki et al. (1995) found the DOE in 4-year-old children but not in 5-year-old children when training a two-comparison symbolic-MTS. Likewise, in the first experiment by Estevez, Fuentes, Mari-Beffa, Gonzalez, and Alvarez (2001), the DOE was demonstrated in children between 4 and 7 years, but not in children 8 years and older. Evidence of age as a moderator of the DOE can also be found in Lopez-Crespo et al. (2009), which compared the DOP in younger adults (undergraduate students) and older adults (around 60 years), and found that DOP enhanced performance on delayed-MTS only in the older group. Further research is needed to explore possible age-related effects of the DOP and the behavioral processes that may be associated with age.

The Differential Outcomes Effect in Clinical Populations

The subgroup analysis of clinical populations showed a large positive effect of the DOP on performance accuracy pooled across studies measuring overall accuracy, terminal accuracy, and transfer. The effect was comparatively greater for clinical populations than for healthy adults and typically developing children when these were used as controls in the same studies. The literature suggests that the DOP could optimize the acquisition of conditional discriminations by improving accuracy in various clinical populations, including children with autism, adults and children with Down syndrome, and individuals with memory impairments (such as age-related memory deficits, Alzheimer’s disease, and sleep deprivation).

As part of our systematic review, we identified eight SSED studies conducted with clinical populations. Although we were unable to include these studies in the meta-analysis, they generally confirmed the results of the subgroup analysis. Specifically, most participants demonstrated improved performance under the DOP, particularly for terminal accuracy and transfer (see Supporting Information). When sessions or trials to mastery were recorded, most participants acquired conditional discrimination in fewer sessions under DOP than

NDP. However, differences in this measure of performance were often small, suggesting that sessions to mastery might be less sensitive to the DOE than other outcomes.

Discrimination *Difficulty* and the Differential Outcomes Effect

We observed larger effect sizes on overall accuracy among experiments implementing symbolic MTS and delayed MTS compared with studies using identity MTS. Both symbolic and delayed MTS increase the *difficulty* of conditional discrimination. Specifically, symbolic MTS features purely arbitrary sample-comparison relations, whereas delayed MTS requires participants to remember the sample. Some studies have increased difficulty within delayed MTS by increasing the number of stimuli or adding a distracting task, such as reciting a series of numbers backwards, to deplete the stimulus control exerted by the sample. For example, Estevez et al. (2001) only demonstrated the DOE in children after increasing the number of comparison stimuli and increasing the similarity between stimuli. Moreover, Plaza, Estevez, Lopez-Crespo, and Fuentes (2011) attained a more distinct DOE by asking participants to recite a number sequence backwards during the sample-comparison time delay, thereby preventing rehearsal. The DOP is said to overcome difficulty by providing an additional and more salient cue to correct responding even on occasions when there is a delay between sample presentation and response (Jones & White, 1994).

We found larger effects for transfer relative to overall and terminal accuracy. This finding may also be in line with the difficulty hypothesis. Specifically, transfer tests influence discrimination difficulty by way of interference from previously acquired discriminations. By introducing novel stimuli following initial training, performance returns to baseline levels, allowing the DOE to be revealed. These findings suggest that the DOP may be affected by task difficulty, with larger effects correlated with more difficult conditional discrimination tasks, such as delayed MTS and transfer.

In addition, we found larger effects for populations with clinical features, such as memory impairments and intellectual disability, than for typically developing children and healthy controls. As already noted, children are often less accurate in acquiring conditional discrimination than adults. Similarly, clinical populations, particularly those with cognitive and memory impairments, may perform poorly in conditional discrimination tasks (Green, 2001; Saunders & Spradlin, 1989). The effect of the DOP may be characterized by a multi-dimensional construct labeled here as difficulty, which could be operationalized in terms of the complexity of the sample-comparison relation. Goeters at el. (1992) suggested that DOE is most prominent in the animal literature when stimulus control is more difficult to establish-- for example, when the stimuli are complex or difficult to discriminate, or when a large number of comparison stimuli are used. Goeters et al. noted that the DOE is more apparent in animals under delayed MTS and match-to-position experimental preparations. The current analysis is consistent with the Goeters et al. hypothesis for animals.

Practical Implications

The difficulty hypothesis has a number of potential clinical implications. For example, the DOP may be an effective intervention when subjects fail to acquire conditional discriminations through standard means or stimuli are especially complex, and therefore more difficult to master. Similarly, when the stimulus set is larger, the discriminative properties of the reinforcer may become more salient. Thus, DOP may have a greater practical value for the acquisition of relatively advanced discrimination skills.

Using the DOP for transfer could be effective in facilitating generalization of already acquired responses to novel stimuli and establishing a stimulus class. By pairing novel stimuli with the same outcomes used in training, the outcomes could cue the correct response and facilitate transfer. Dube McIlvane, Maguire, Mackay, and Stoddard(1989) showed that the outcomes used in the DOP become part of a stimulus class consisting of the stimulus, response, and outcome. Pairing stimuli with the differential outcomes used in training can therefore be used to add new samples to the stimulus class (see also Dube et al., 1987, Experiment 2). A number of studies have demonstrated that the DOP can be used to facilitate equivalence with individuals with intellectual disabilities and autism spectrum disorder (Joseph et al., 1997; Varella & De Souza, 2015). For example, Joseph et al. showed enhanced accuracy in equivalence testing following conditional discrimination training under DOP relative to NDP. Moreover, arranging class-specific reinforcers (analogous to arranging a DOP) is sufficient to mediate class merging and expansion (e.g., Dube et al., 1989; Maki et al., 1995; Minster et al., 2006). This area is in need of further research, especially because standard conditional discrimination procedures do not reliably result in equivalence and generalization in individuals with limited language abilities (O'Donnell & Saunders, 2003).

In several studies included in our analysis, tokens were used to reinforce correct responses. Tokens may differ both in color and in the rewards that they are exchanged for. However, in Estevez, Overmier, and Fuentes (2003b) the authors demonstrated that the DOP can be effective even when these tokens are exchanged for the same primary reinforcers. This is a useful adaptation, provided that tokens are effective reinforcers for the participants' behavior.

It has been noted that the DOP is more onerous to carry out than conditional discrimination with a common outcome, due to the need for multiple highly preferred reinforcers (Chong, 2005). For example, Litt and Schreibman (1981) reported that the DOP is only effective if reinforcers are equally preferred, which may restrict the pool of effective reinforcers. Alternatively, it may be necessary to implement a black-out period for incorrect responses to deter over-selection of the stimulus associated with the preferred reinforcer on occasions when the reinforcers have different magnitudes. While the DOP may require more preparation than typical conditional discrimination procedures, it may result in socially important effects on acquisition, maintenance, and generalization in the context of extended instruction (e.g., early intensive behavioral intervention for children with autism spectrum disorder).

One important extension for future research into the DOE may be the inclusion of clinically important outcomes such as trials to mastery. Specifically, trials to mastery may be a preferred acquisition outcome when comparing the efficiency of the DOP with other forms of intervention.

Limitations

Our analysis was limited to group studies. The SSED literature made up only a small proportion of the studies examined in our review. Only three of the nine SSED experiments captured by our search provided session-by-session datasets. These were provided by Chong (2004, Experiments 1 and 2) and Addison (2006). The remaining six studies reported aggregated outcomes. Without session-by-session datasets, a full meta-analysis is not possible. Some of the factors that may affect the DOP, such as sensory modality and dependent variables, were not directly comparable across the SSED and group studies. Additionally, several of the group studies were authored by a relatively small group of scientists and may therefore benefit from further replications. Finally, our moderator analysis suggested that the methodological standards in this literature may have an impact on the DOE. Specifically, studies with relatively lower methodological standards had a tendency to over-estimate overall and terminal accuracy. As the body of literature on the DOE becomes more extensive, it may be possible to determine whether this bias can be attributed to a discrete factor. A recent study has shown that quality standards such as sample size and randomization tend to reduce effect sizes when evaluating educational programs (Cheung & Slavin, 2016), although these standards were not a concern in our study.

Conclusion

The current analysis showed a medium-to-large effect of DOP on different accuracy measures (overall, terminal, and transfer), providing strong evidence that the procedure is effective in humans. While the primary literature base for this conclusion comes from group experiments, SSED studies served as a source of systematic replication. According to this literature, most participants perform better under the DOP. Single-subject analyses also revealed individual differences in the effectiveness of DOP that warrants further research.

The subgroup analyses suggest that age mediated DOE. The potential age effect, while not identifiable with any particular behavioral process, may suggest that those with more limited discriminative repertoires are more likely to benefit from the exposure to differential outcomes. When joined with evidence from the moderator analysis, this finding suggests that DOP may be mediated by task difficulty. Thus, DOP may be of special practical value to participants demonstrating poor conditional discrimination, such as individuals with intellectual disabilities and pervasive developmental disorders. The SSED literature and clinical group data demonstrate the potential of the DOP in teaching individuals with cognitive and memory impairments. For a number of participants, the procedure generated significantly more accurate performance and decreased the number of sessions required to reach mastery. The clinical and methodological diversity of the studies synthesized here support the generality of the DOE. The diversity of this literature should provide the impetus for future experimental analyses that will help to establish the individual and procedural factors that would optimize the DOE.

Despite the need for further research, it does not seem too soon to recommend that DOPs are arranged whenever practicable. This review suggests that it is reasonable to expect benefits in various aspects of learning that should outweigh the inconvenience of using a slightly more complex procedure. DOPs are also more typical of naturally occurring contingencies, in which different kinds responses naturally lead to different kinds of reinforcers. That is, the natural environment usually arranges differential outcomes. In this light, the arrangement of the same reinforcer for different responses in most experimental and applied research seems both to lack ecological validity and, on the basis of the results we have reported here, is likely to have a deleterious effect on discrimination performance. Perhaps we should begin to speak of a suppressive *same*-outcomes effect that artificially reduces performance relative to what we might expect from the more natural contingencies, rather than a differential-outcomes effect that enhances performance.

References

Addison, L. (2006). *An examination of the differential outcomes effect when teach discrimination to children with autism and other developmental disabilities* (Doctoral dissertation). Retrieved from Louisiana State University Digital Commons (Accession no. etd-06022006-184225).

Baldwin, R., Chelonis, J., Prunty, P., & Paule, M. (2012). The use of an incremental repeated acquisition task to assess learning in children. *Behavior Processes, 91*, 104-114. 10.1016/j.beproc.2012.06.004

Borenstein, M., Hedges, L.V., Higgins, J. & Rothstein, H. (2009). *Introduction to Meta-Analysis.* Chichester, UK: Wiley.

Borenstein, M., Hedges, L. V., & Rothstein, H. (2007). *Meta-Analysis Fixed effect vs. random effects* [pdf version]. Retrieved from: https://www.meta-analysis.com/downloads/M-a_f_e_v_r_e_sv.pdf

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283-292. doi:10.3102/0013189X16656615

Chong, I. (2004). *Assessing the differential outcomes procedure with children diagnosed with autism* (Doctoral dissertation). Western Michigan University Dissertations (Accession no. 1086).

Chong, I. & Carr, J. (2010). Failure to demonstrate the differential outcomes effect in children with autism. *Behavioral Interventions, 25,* 339-348. doi:10.1002/bin.318

de Wit, S., Corlett, P., Aitken, M., Dickinson, A., & Fletcher, P. (2009). Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *Journal of Neuroscience, 29,* 11330-11338. doi:10.1523/JNEUROSCI.1639-09.2009

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177-188. doi:10.1016/0197-2456(86)90046-2

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology Community Health, 52*, 377-384. doi:10.1136/jech.52.6.377

Dube, W. V., McIlvane, W. J., Mackay, H. A., & Stoddard, L. T. (1987). Stimulus class membership established via stimulus-reinforcer relations. *Journal of the Experimental Analysis of Behavior, 47*, 159-175. doi:10.1901/jeab.1987.47-159

Dube, W., McIlvane, W., Maguire, R., Mackay, H., & Stoddard, L. (1989). Stimulus class formation and stimulus-reinforcer relations. *Journal of the Experimental Analysis of Behavior, 51*, 65-76. doi:10.1901/jeab.1989.51-65

Easton, A. (2004). Differential reward outcome learning in adult humans. *Behavior Brain Research, 154*, 165-169. doi:10.1016/j.bbr.2004.02.023

Easton, A., Child, S., & Lopez-Crespo, G. (2011). Differential outcomes aid the formation of categorical relationships between stimuli. *Behavior Brain Research, 222*, 270-273. doi:10.1016/j.bbr.2004.02.023

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634. doi:10.1136/bmj.315.7109.629

Esteban, L., Plaza, V., Lopez-Crespo, G., Vivas, A., & Estevez, A. (2014). Differential outcomes training improves face recognition memory in children and adults with Down syndrome. *Research and Developmental Disabilities, 35*, 1384-1392. doi:10.1016/j.ridd.2014.03.031

Esteban, L. Vivas, A., & Estevez, A. (2014). Visual recognition memory enhancement in children through differential outcomes. *Acta Psychologica, 150*, 146-152. doi:10.1016/j.actpsy.2014.05.005

Estevez, A., Fuentes, L., Mari-Beffa, P., Gonzalez, C., & Alvarez, D. (2001). The differential outcome effect as a useful tool to improve conditional discrimination learning in children. *Learning and motivation, 32*, 48-64. doi:10.1006/lmot.2000.1060

Estevez, A., & Fuentes, L. (2003). Differential outcomes effect in four-year-old children. *Psicológica, 24*, 159-167.

Estevez, A., Fuentes, L., Overmier, J.B., & Gonzalez, C. (2003a). Differential outcomes effect in children and adults with Down syndrome. *American Journal of Mental Deficiency, 108,* 108-116. doi:10.1352/0895-8017(2003)108%3C0108:DOEICA%3E2.0.CO;2

Estevez, A., Overmier, B., & Fuentes, L. (2003b). Differential outcomes effect in children: Demonstration and mechanism. *Learning and motivation*, *34*, 148-167. doi:10.1016/S0023-9690(02)00510-6

Estevez, A., Vivas, A., Alonso, D., Mari-Beffa, P., Fuentes, L., & Overmier, J. B. (2007). Enhancing challenged students' recognition of mathematical relations through differential outcomes training. *The Quarterly Journal of Experimental Psychology, 60*, 571-580. doi:10.1080/17470210600820039

Flores, C., Ortega, D., Reyes, K., Mateos, R., Villanueva, S., & Amaya, A. (2005). Sample-comparative partial and total reversals in matching-to-sample with differential and no differential outcomes. *Universitas Psychologica, 4,* 43-47.

Goeters, S., Blakely, E., & Poling, A. (1992). The differential outcomes effect. *Psychological Record*, *42,* 389-409. doi:10.1007/BF03399609

Green, G. (2001). Behavior analytic instruction for learners with autism advances in stimulus control technology. *Focus on Autism and Other Developmental Disabilities*, *16*(2), 72-85. doi:10.1177/108835760101600203

Harbord, R. M., & Higgins, J. P. T. (2008). Meta-regression in Stata. *Stata Journal, 8*(4), 493-519. Retrieved from http://www.stata-journal.com/sjpdf.html?articlenum=sbe23_1

Hedges, L. V. (1981) Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128. doi:10.2307/1164588

Higgins, J., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions, Version 5.1.0.* The Cochrane Collaboration. Available from www.handbook.cochrane.org

Higgins J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558. doi:10.1002/sim.1186

Hochhalter, A., Sweeney, W., Bakke, B., Holub, R., & Overmier, J.B. (2001). Improving face recognition in alcohol dementia. *Clinical Gerontologist, 22*, 3-18. doi:10.1300/J018v22n02_02

Joseph, B., Overmier, J. B., & Thompson, T. (1997). Food- and non-food-related differential outcomes in equivalence learning by adults with Prader-Willi syndrome. *American Journal of Mental Retardation, 101*(4), 374-386.

Jones, B. M., & White, K. (1994). An investigation of the differential-outcomes effect within sessions. *Journal of the Experimental Analysis of Behavior, 61*, 389-406. 10.1901/jeab.1994.61-389

Kratochwill, T., Hitchcock, J., Horner, R., Levin, J., Odom, S., Rindskopf, D., & Shadish, W. (2010). *Single-case designs technical documentation*. What Works Clearinghouse, Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Document/229

Legge, E., & Spetch, M. (2009). The differential outcomes effect (DOE) in spatial localization: An investigation with adults. *Learning and Motivation, 40*, 313-328. 10.1016/j.lmot.2009.03.002

Litt, M. D., & Schreibman, L. (1981). Stimulus-specific reinforcement in the acquisition of receptive labels. *Analysis and Intervention in Developmental Disabilities, 1*, 171-186.

López-Crespo, G., Plaza, V., Fuentes, L., & Estevez, A. (2009). Improvement of age-related memory deficits by differential outcomes. *International Psychogeriatrics, 21*, 503-510. doi:10.1017/S1041610209008576

López-Crespo, G., Teresa Daza, M., & Mendez-Lopez, M. (2012). Visual working memory in deaf children with diverse communication modes: Improvement by differential outcomes. *Research in Developmental Disabilities, 33*, 362-368. doi:10.1016/j.ridd.2011.10.022

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

Maki, P., Overmier, J. B., Delos, S., & Gutmann, A. J. (1995). Expectancies as factors influencing conditional discrimination performance of children. *The Psychological Record, 45*, 45-71.

Malanga, P., & Poling, A. (1992). Letter recognition by adults with mental handicaps: Improving performance through differential outcomes. *Mental Retardation & Learning Disability Bulletin, 20*, 39-48.

Martella, D., Plaza, V., Estevez, A., Castillo, A., & Fuentes, L. (2012). Minimizing sleep deprivation effects in healthy adults by differential outcomes. *Acta Pscyhologica, 139*, 391-396. doi:10.1016/j.actpsy.2011.12.013

Martinez, L., Flores, P., Gonzalez-Salinas, C., Fuentes, L., & Estevez, A. (2013). The effects of differential outcomes and different types of consequential stimuli on 7-year-old children's discriminative learning and memory. *Animal Learning & Behavior, 41*, 298-308. doi:10.1016/j.ridd.2011.08.022

Martinez, L., Mari-Beffa, P., Roldan-Tapia, D., Ramosi-Lizana, J., Fuentes, L., & Estevez, A. (2012). Training with differential outcomes enhances discriminative learning and visuospatial recognition memory in children born prematurely. *Research in Developmental Disabilities, 33*, 76-84. doi:10.1016/j.ridd.2011.08.022

Miller, O., Waugh, K., & Chambers, K. (2002). Differential outcomes effect: Increased accuracy in adults learning Kanji with stimulus specific rewards. *The Psychological Record, 52*, 315-324. doi:10.1007/BF03395433

Minster, S.T., Jones, M., Elliffe, D., & Muthukumaraswamy, S.D. (2006). Stimulus equivalence: Testing Sidman's (2000) theory. *Journal of the Experimental Analysis of Behavior, 85*, 371-391. doi:10.1901/jeab.2006.15-05

Mok, L., Estevez, A., & Overmier, J.B. (2010). Unique outcome expectations as a training and pedagogical tool. *The Psychological Record, 60*, 227-248. doi:10.1007/BF03395705

Mok, L., Thomas, K., Lungu, O., & Overmier, J. B. (2009). Neural correlates of cue-unique outcome expectations under differential outcomes training: An fMRI study. *Brain Research, 1265*, 111-127. doi:10.1016/j.brainres.2008.12.072

O'Donnell, J., & Saunders, K. (2003). Equivalence relations in individuals with language limitations and mental retardation. *Journal of the Experimental Analysis of Behavior, 80*, 131-157. doi:10.1901/jeab.2003.80-131

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

Peterson, G. (1984). How expectancies guide behaviour. In H.L. Roitblat, T.G. Bever, & H.S. Terrace (Eds.), *Animal Cognition* (pp. 135-148). Hillsdale, NJ: Erlbaum.

Peterson, G., & Trapold, M. (1982). Expectancy mediation of concurrent conditional discrimination. *The American Journal of Psychology, 95*, 571-580. doi:10.2307/1422188

Plaza, V., Esteban, L., Estevez, A., & Fuentes, L. (2013). The differential outcomes procedure improves face recognition irrespective of their emotional valence. *Psicologica, 34*, 79-95.

Plaza, V., Estevez, A., Lopez-Crespo, G., & Fuentes, L. (2011). Enhancing recognition memory in adults through differential outcomes. *Acta Psychologica, 136*, 129-136. doi:10.1016/j.actpsy.2010.11.001

Plaza, V., Lopez-Crespo, G., Antunez, C., Fuentes, L., & Estevez, A. (2012). Improving delayed face recognition in Alzheimer's disease by differential outcomes. *Neuropsychology, 26,* 483-489. doi:10.1037/a0028485

Ruge, H., Krebs, R. M., & Wolfensteller, U. (2012). Early markers of ongoing action-effect learning. *Frontiers in Psychology, 3*, 522. doi:10.3389/fpsyg.2012.00522

Ruge, H., & Wolfensteller, U. (2013). Functional integration processes underlying the instruction-based learning of novel goal-directed behaviors. *Neuroimage, 68*, 162-172. doi:10.1016/j.neuroimage.2012.12.003

Saunders, R. R., & Sailor, W. (1979). A comparison of three strategies of reinforcement on two-choice learning problems with severely retarded children. *Research and Practice for Persons with Severe Disabilities, 4*(4), 323-333. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/154079697900400401

Saunders, K. J., & Spradlin, J. E. (1989). Conditional discrimination in mentally retarded adults: The development of generalized skills. *Journal of the Experimental Analysis of Behavior, 54*, 239-250. doi:10.1901/jeab.1990.54-239

Schmidtke, K. A. (2010). *The differential outcome effect with typical adult humans* (Doctoral dissertation). Retrieved from Auburn University Electronic Theses and Dissertations (Accession no. 2009-07-02T17:55:45Z).

Schmucker, C. M., Blümle, A., Schell, L. K., Schwarzer, G., Oeller, P., Cabrera, L., . . . Meerpohl, J. J. (2017). Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. *PLoS ONE, 12*, e0176210. doi:10.1371/journal.pone.0176210

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior*, *74*, 127-146. doi:10.1901/jeab.2000.74-127

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37*, 5-22. doi:10.1901/jeab.1982.37-5

Sterne, J., Harris, R., Harbord, R., & Steichen, T. (2009). *Meta-analysis in Stata: An updated collection from the Stata Journal.* College Station, Texas: Stata Press.

Trapold, M. (1970). Are expectancies based upon different reinforcing events discriminably different? *Learning and Motivation, 1*, 129-140. doi:10.1016/0023-9690(70)90079-2

Urcuioli, P. (1990). Differential outcomes and many-to-one: Effects of correlation with correct choice. *Animal Learning & Behavior, 18*, 410-422. doi:10.3758/BF03205323

Urcuioli, P. (2005). Behavioral and associative effects of differential outcomes in discrimination learning. *Learning and Behavior, 33*, 1-21. doi:10.3758/BF03196047

Varella, A. A., & De Souza, D.G. (2015). Using class-specified compound consequences to teach dictated and printed letter relations to a child with autism. *Journal of Applied Behavior Analysis, 48*, 675-679. doi:10.1002/jaba.224.

Virués-Ortega J. (2010). Applied behavior analytic intervention for autism in early childhood: meta-analysis, meta-regression and dose-response meta-analysis of multiple outcomes. *Clinical Psychology Review, 30*, 387-399. doi:10.1016/j.cpr.2010.01.008.

Received March 11, 2017

Final acceptance September 6, 2018

Action Editor, Einar Ingvarsson

Table 1
Characteristics of Studies

Variable	Systematic Review		Meta-Analysis	
	n	%	*n*	%
Distinct experiments (publications)	60 (35)		43 (26)	
Age				
Less than 6	14	23.3	10	23.3
6 to 12	12	20.0	10	23.3
12 to 18	0	0	0	0
Over 18	34	56.7	24	55.8
Over 60	3	5.0	2	4.7
Clinical conditions				
Autism spectrum disorder	4		0	
Intellectual Disability	5		2	
Memory impairments	4		3	
Other	3		2	
Termination criteria				
40 or fewer trials	11	18.3	6	25.6
41 to 72 trials	20	33.3	18	41.9
73 to 135 trials	19	31.7	16	37.2
Mastery	10	16.7	3	7.0
Training format				
Conditional discrimination (unspecified)	16	26.7	10	23.3
Identity MTS	6	10.0	5	11.6
Symbolic MTS	29	48.3	25	58.1
Other	9	15.0	3	7.0
No Delay	19	31.7	6	14.0
With Delay	39	65.0	37	86.0
Reinforcement type				
Token reinforcement	15	25.0	13	30.2
Other	45	75.0	30	69.8
Control				
Variable outcomes	52	86.7	39	90.7
Common outcome	10	16.7	6	14.0
Measures of behavior				
Overall accuracy	42	70.0	36	83.7
Terminal or test accuracy	20	33.3	12	46.5
Transfer	9	15.0	7	20.9
Latency	18	30.0	17	39.5
Errors	3	5.0	3	7.0
Mastery	6	10.0	3	7.

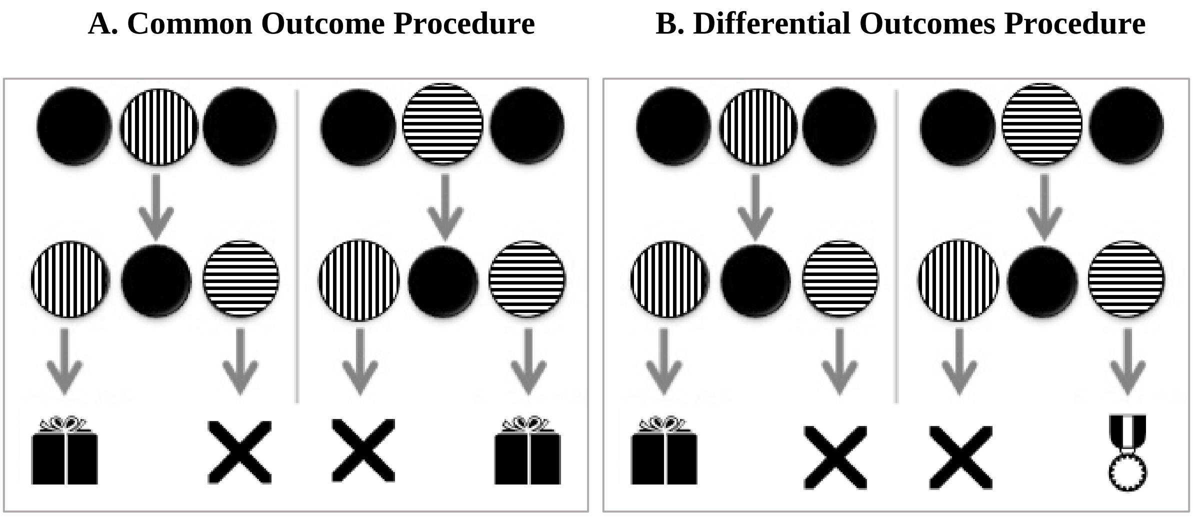
Note. Unless otherwise indicated *n* refers to the number of distinct experiments. One experiment could be assigned to more than one category.

Table 2

Moderator and Subgroup Analyses

	n	Overall Accuracy	*n*	Terminal Accuracy	*n*	Transfer	*n*	Latency
		ES [95% CI]		ES [95% CI]		ES [95% CI]		ES [95% CI]
Age, years		*ns*		*ns*		*ns*		*ns*
Child (≤17)	20	0.71 [0.27, 1.15]	3	1.09 [-0.05, 2.23]	7		1	
Adult (>17)	20	0.44 [0.00, 0.87]	8	0.75 [0.08, 1.41]	0		19	-0.33 [-0.78, 0.12]
Population		*ns*				*ns*		*p* = 0.065
Clinical	7	0.99 [0.27, 1.72]	11		1		1	-
Non-clinical	33	0.48 [0.13, 0.82]	0		6		19	-0.23 [-0.68, 0.22]
Format #1		*p* = 0.115[1]						*ns*
Conditional	14	0.29 [-0.24, 0.82]	0		0		14	-0.15 [-0.68, 0.38]
Identity MTS	6	0.49 [-0.20, 1.18]	0		0		3	-0.88 [-2.14, 0.38]
Symbolic MTS	20	0.81 [0.36, 1.26]	10		7		3	-0.53 [-1.63, 0.57]
Format #2		*ns*		*ns*				*ns*
No-Delay	5	0.42 [-0.47, 1.30]	2	0.60 [-0.80, 2.00]	0		8	-0.37 [-1.07, 0.34]
Delayed MTS	35	0.59 [0.26, 0.92]	9	1.03 [0.35, 1.71]	7		12	-0.29 [-0.85, 0.27]
Reinforcer type		*p* = 0.032		*ns*				
Token	10	1.23 [0.59, 1.88]	2	1.28 [-0.09, 2.64]	0		0	
Other	30	0.40 [0.05, 0.75]	9	0.75 [0.08, 1.41]	7		20	
Duration		*ns*		*ns*		*p* = 0.111		*ns*
≤72 trials[2]	14	0.77 [0.24, 1.30]	4	1.03 [0.04, 2.02]	5	1.20 [0.39, 2.01]	5	-0.42[-1.30, 0.46]
>72 trials	26	0.47 [0.09, 0.85]	7	0.74 [0.03, 1.44]	2	3.14 [0.62, 5.66]	15	-0.29 [-0.80, 0.22]
Quality		*p = 0.042*		*p = 0.042*		*ns*		*ns*
<10[2]	17	0.79 [0.31, 1.27]	6	1.01 [0.14, 1.88]	6		9	-0.44 [-1.09, 0.20]
≥10	23	0.42 [0.02, 0.82]	5	0.81 [-0.08, 1.71]	0		11	-0.22 [-0.81, 0.38]

Notes. Sensitivity and moderator analyses left blank due to insufficient studies to perform the analysis. MTS = Matching to sample; *n* = number of studies; *ns* = non-significant moderator analysis. 1. All *p* values pertain to the moderator analyses coefficients reported in the text. 2. Median used as cut-off point.



A. Common Outcome Procedure

B. Differential Outcomes Procedure

Figure 1. Example of matching-to-sample task using a nondifferential outcomes procedure (common outcome) and the differential outcomes procedure.

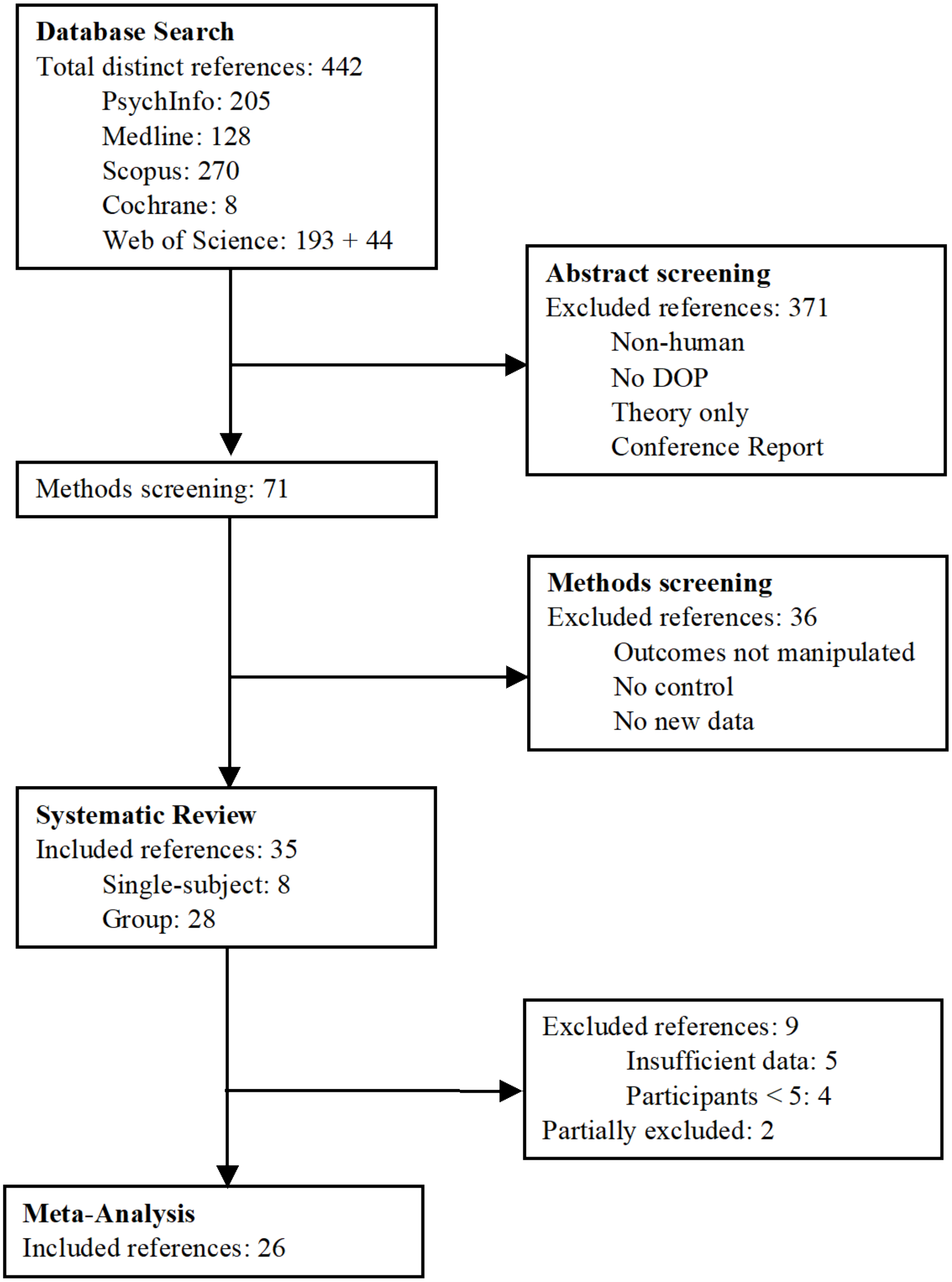


Figure 2. Selection of references.

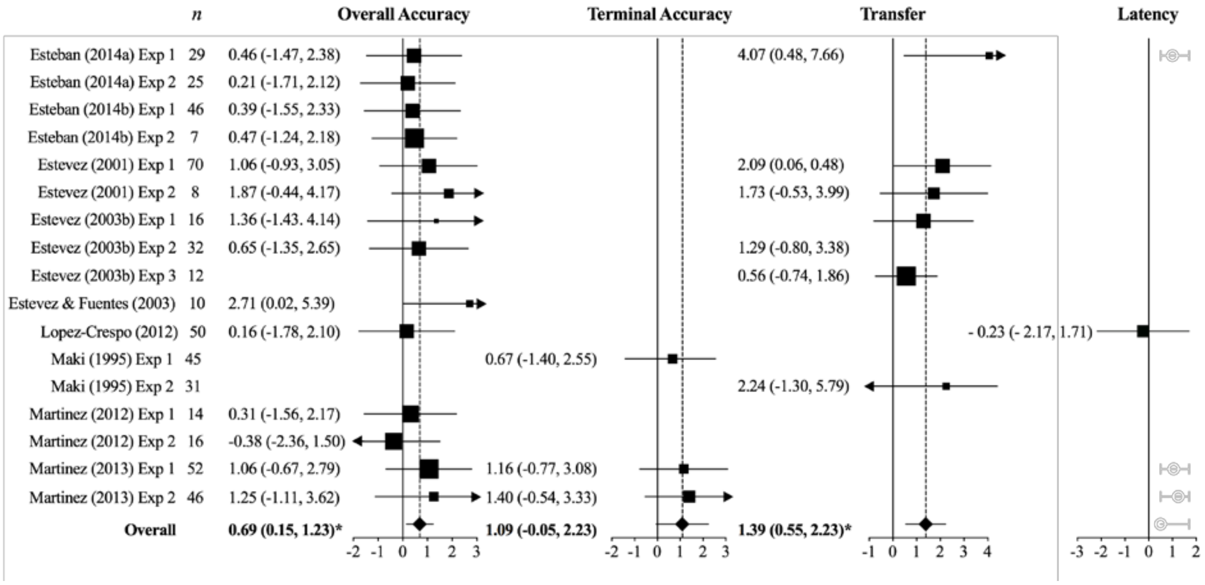


Figure 3. Weighted mean effect size and 95% confidence interval of differential reinforcers in studies with children. Symbol size is proportional to the study's contribution to the overall effect. Horizontal lines represent confidence intervals. Confidence intervals extending beyond the range of the graph are marked with an arrow. Dotted vertical lines denote the overall effect size. Solid vertical lines mark the non-effect level. All overall effects are significant (*p* < 0.5).

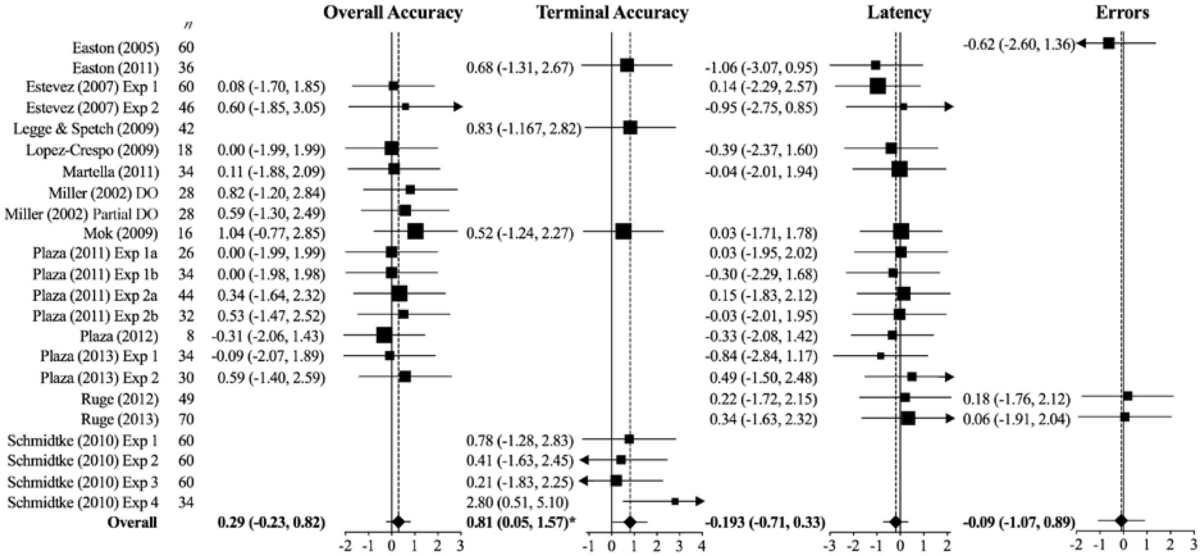
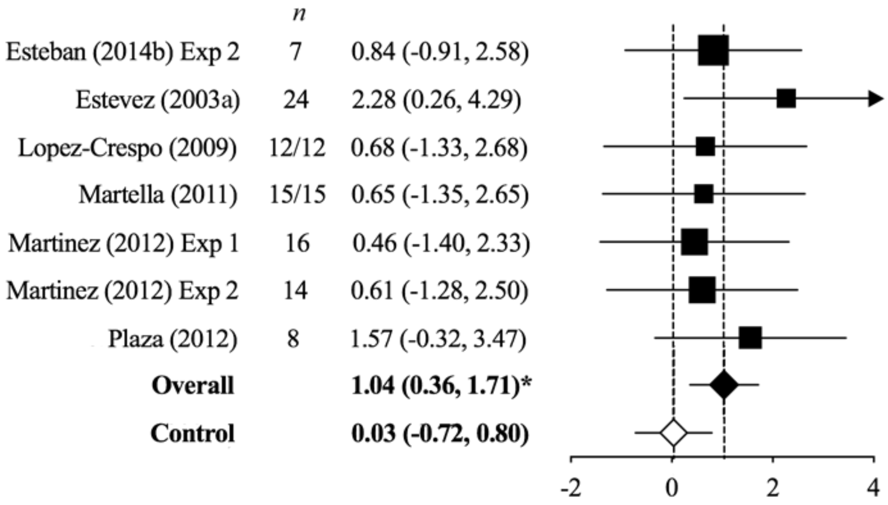


Figure 4. Weighted mean effect size and 95% confidence interval of differential reinforcers in studies with adult participants. Symbol size is proportional to the study's contribution to the overall effect. Miller et al. partial differential outcome within-subjects condition (partial DO) was not included in the meta-analysis. * *p* < 0.5 (overall effect).



	n	
Esteban (2014b) Exp 2	7	0.84 (-0.91, 2.58)
Estevez (2003a)	24	2.28 (0.26, 4.29)
Lopez-Crespo (2009)	12/12	0.68 (-1.33, 2.68)
Martella (2011)	15/15	0.65 (-1.35, 2.65)
Martinez (2012) Exp 1	16	0.46 (-1.40, 2.33)
Martinez (2012) Exp 2	14	0.61 (-1.28, 2.50)
Plaza (2012)	8	1.57 (-0.32, 3.47)
Overall		**1.04 (0.36, 1.71)***
Control		**0.03 (-0.72, 0.80)**

Figure 5. Weighted mean effect size and 95% confidence interval of differential reinforcers over performance accuracy (percentage of correct) in studies with clinical population. Symbol size is proportional to the study’s contribution to the overall effect. * *p* < 0.5 (overall effect).