Behavior Analysis in Practice
https://doi.org/10.1007/s40617-025-01140-2

DISCUSSION AND REVIEW PAPER

Matched Sensory Stimulation to Reduce Automatically Maintained Challenging Behavior: A Systematic Review and Meta-Analysis

Tyler-Curtis C. Elliott[1] · Chloe N. Wise[1] · Jessica N. Torelli[1] · Kevin M. Ayres[1] · Joel E. Ringdahl[1]

Received: 16 December 2023 / Accepted: 21 November 2025

Abstract
Automatically reinforced challenging behavior presents unique intervention challenges because the behavior produces its own reinforcer. Matched stimulation (Piazza et al., *Journal of Applied Behavior Analysis*, *31*, 165–189, 1998) is a method of providing access to the putative sensory consequences that maintain challenging behavior through interaction with an alternative stimulus (e.g., toys, items). This approach may reduce motivation to engage in automatically reinforced challenging behavior, resulting in response reduction. In this article, we conducted a systematic literature review and meta-analysis of 32 studies that examined matched stimulation to reduce automatically maintained challenging behavior. We descriptively coded each article at the participant level to extract information about behaviors, participants, assessments, and intervention variables and determine how the effects of matched stimulation have been evaluated in the literature. We then conducted a random effects meta-analysis to estimate the average effects of matched stimulation. We found that matched stimuli produced a statistically significant decrease in challenging behavior when compared to a control condition and when compared to unmatched stimuli. Last, we conducted a moderator analysis to explore how assessment and treatment characteristics influenced the magnitude of effect, which revealed that functional behavior assessment type and number of matched stimuli moderated intervention effects, whereas the use of competing stimulus assessments and preference assessments did not moderate the effectiveness of matched stimulation.

Keywords Matched stimulation · Automatic reinforcement · Noncontingent reinforcement

The advent of functional analysis as an assessment methodology (e.g., Iwata et al., 1982/1994; Northup et al., 1991; Vollmer et al., 1995) to identify variables responsible for maintaining challenging behavior[1] significantly increased the success of interventions to decrease these behaviors (Campbell, 2003). Functional analyses of challenging behavior yield one of three outcomes: (a) The results are undifferentiated due to low response rates or outcomes that are not replicable; (b) the behavior is reinforced by socially mediated reinforcers (e.g., attention and escape); or (c) the behavior is reinforced by nonsocially mediated reinforcers (i.e., reinforcers that do not require another individual to control the reinforcer delivery; Vaughan & Michael, 1982). When a functional analysis of challenging behavior suggests that nonsocially mediated reinforcers maintain a behavior, that behavior is often referred to as automatically reinforced.

Automatically reinforced challenging behavior is often associated with specific barriers to intervention (LeBlanc

✉ Tyler-Curtis C. Elliott
tce62651@uga.edu

Chloe N. Wise
chloewise99@yahoo.com

Jessica N. Torelli
jessica.torelli@uga.edu

Kevin M. Ayres
kayres@uga.edu

Joel E. Ringdahl
ringdahl@uga.edu

[1] Center for Autism and Behavioral Education Research, University of Georgia, Rivers Crossing, 850 College Station Road Bldg. 2, Athens, GA 30605, USA

[1] We understand that the phrasing “challenging behavior” has a controversial connotation. Readers who see the words “challenging behavior” should consider the social validity of the target behavior and ask themselves “Challenging for whom? The child or the researcher?” In this article, we refer to various behaviors that interfere with safety, independent living, least restrictive environment, and community access for the person with developmental disabilities. Rather than include this phrasing each time we reference these types of behaviors, we elected to use the phrase challenging behavior.

Published online: 24 February 2026

et al., 2000). For example, the reinforcer and its specific properties (e.g., magnitude) may not be identifiable (Rooker et al., 2018; Vollmer, 1994). In addition, even if one identifies the properties of reinforcement relevant to automatically reinforced behavior, they may not be able to control its delivery (Vollmer, 1994). For example, functional communication training often includes extinction of the challenging behavior and reinforcement of the alternative behavior with a functionally matched reinforcer (Tiger et al., 2008). If the behavior produces its own reinforcer (e.g., biting one's own tongue for the feeling it produces), the researcher can neither extinguish this response-reinforcer relation (e.g., block the internal feeling produced or removed by biting the tongue), nor can they deliver this tongue input contingent on the occurrence of the alternative behavior (Vaughan & Michael, 1982).

Despite these inherent difficulties in designing interventions for automatically reinforced challenging behavior, the existing literature details numerous examples of successful intervention approaches including antecedent-based interventions (e.g. noncontingent reinforcement [NCR]; Ringdahl et al., 1997), reinforcement-based procedures (e.g., differential reinforcement of other behavior [DRO]; Nuernberger et al., 2013), and interventions that include default components (i.e., those that are likely to be effective regardless of the maintaining function; e.g., Berg et al., 2016; Roscoe et al., 2013).

Matched Stimulation

One common reinforcement-based method for reducing automatically reinforced challenging behavior is matched stimulation. The term matched stimulation implies that a stimulus used in intervention provides the same sensory stimulation as the automatically reinforced challenging behavior and can thereby replace the reinforcers produced by that behavior, reducing its occurrence. For example, Piazza et al. (2000) identified shaving cream as a matched stimulus for saliva because noncontingent access to the shaving cream led to a reduction in saliva play exhibited by the participant and was hypothesized to match the sensory reinforcer provided by saliva play. In addition, some researchers have evaluated matched stimuli prior to the emergence of the term "matched stimuli" in the literature. For example, Favell et al. (1982) discuss potential "sensory match" between their target behavior and alternative "sensory activities" (e.g., visual toys). Researchers have since experimentally demonstrated cases in which matched stimulation successfully reduced challenging behavior (e.g., Piazza et al., 2000), and instances in which it did not (e.g., Lanovaz et al., 2009). In addition, researchers have provided examples of interventions that incorporated unmatched stimulation producing behavior reductions equal to or better than those produced by matched stimulation (e.g., Ahearn et al., 2005).

Given that the underlying stimulus changes produced by automatically reinforced behaviors can be private events, researchers can only assume (based on the features of the stimulus and the client response or based on an assessment) that the selected stimulus will provide the same sensory stimulation as the challenging behavior. Even a successful reduction in challenging behavior does not guarantee that matched stimulation is the underlying mechanism responsible for the change or lack thereof (Lanovaz et al., 2010). The positive effects of the matched stimulus could be simply because the stimulus competes as a reinforcer when placed in a concurrent operant arrangement. One method for evaluating if a stimulus serves as a completing reinforcer is to conduct a competing stimulus assessment (CSA). This assessment involves measuring both challenging behavior and engagement with a stimulus during a single-stimulus, free-operant arrangement (e.g., Ringdahl et al., 1997).

Haddock and Hagopian (2020) conducted a systematic literature review of the effects of stimuli selected using a CSA on reductions in challenging behavior. Although matched stimuli are not a requisite feature of a CSA, these authors coded whether researchers reported stimuli as matched or unmatched. In their analysis, the authors found that nonfood stimuli reduced challenging behavior by 80% or more in 38% of cases in which stimuli were matched and 43% of cases in which stimuli were unmatched, suggesting minimal difference in intervention efficacy between these classifications. However, Haddock and Hagopian also reported that no researchers who labeled stimuli as matched or unmatched provided a process for making this distinction. Thus, the stimuli that researchers labeled as matched may not have functionally served as a matched stimulus. In addition, Haddock and Hagopian only reviewed studies that used a CSA, limiting the sample. In our own analysis of Haddock and Hagopian's data, matched stimuli increased challenging behavior in only 7.69% of cases whereas unmatched stimuli increased challenging behavior in 19.44% of cases. Given these limitations, we cannot yet assume there is no difference in intervention efficacy between matched and unmatched stimuli.

Variations in Matched Stimulus Interventions

Selection of Matched Stimuli

Researchers have taken a few approaches to identifying matched stimuli: (a) inferring based on the sensory class (e.g., Saylor et al., 2012), (b) comparing to unmatched stimuli (e.g., Piazza et al., 1998), (c) assessing motivating operations (MO; e.g., Lanovaz et al., 2010) or (d) running a sensory assessment (e.g., Rincover, 1978). First, some researchers infer items that are likely to provide matched stimulation based on the hypothesized sensory receptors used when interacting with the stimulus (e.g., Higbee et al., 2005) and by the properties of the stimulus itself (e.g., Fisher et al., 1998). For example, poking oneself in the eye does not involve the gustatoreceptor (the sense receptors located in the mouth), thus providing a stimulus that involves the gustatorceptor (e.g., providing a snack) is unlikely to provide matched stimulation for the client (although it may provide competing, unmatched stimulation). As another example, if based on direct observation, a client engages in pica with soft objects more frequently than hard objects (e.g., Piazza et al., 1998), the assumption is that soft stimuli will be more likely to provide matched stimulation. Second, other researchers have gone a step further by directly comparing the hypothesized matched stimuli with hypothesized unmatched stimuli, to validate their hypothesis that the inferred stimulus was matched (e.g., Taylor et al., 2005).

Third, Lanovaz et al. (2010) proposed another method for identifying stimuli as matched: a three-component multiple schedule that begins with a no-interaction condition, then noncontingent access to the matched stimulus, and a return to the no-interaction condition. These researchers argued that even if responding declines in the second component (matched stimulation) when compared to the first component (no interaction), it does not mean that the stimulus is matched. Rather, this reduction could mean that the stimulus just competes with the source of automatic reinforcement (e.g., Carr et al., 2000; Haddock & Hagopian, 2020). They suggest that if the stimulus is functionally matched, the third component will result in lower levels of challenging behavior than the first, because recent access to this stimulus would serve as an abolishing operation. However, if not matched, the stimulus may impose a state of deprivation for the automatic reinforcer, resulting in higher levels of challenging behavior during the third component than the first.

Finally, one other method for determining if a stimulus is functionally matched is to conduct an assessment to isolate the specific properties maintaining the challenging behavior prior to selecting a stimulus. The most common sensory assessment relies on sensory extinction (e.g., Kennedy & Souza, 1995; Rincover, 1978). The researchers block one automatic consequence at a time and see if the target behavior extinguishes. For example, if researchers hypothesize that spinning items is reinforced by the production of auditory stimulation, they may put carpet on the client's desk such that the behavior no longer produces auditory stimulation (Rincover, 1978). If the occurrence of behavior reduces (and the researchers have demonstrated a functional relation), they use that automatic consequence when selecting the matched stimulus. Another variation of this approach involves experimentally comparing stimuli from each sensory modality. For example, Slocum et al. (2021) directly compared several stimuli hypothesized to produce auditory and tactile information and experimentally identified which class of stimuli successfully reduced the challenging behavior.

Although researchers may use any of these methods to identify their stimuli as potentially matched, they may also supplement their decision making with a CSA, preference assessment (PA), or both. For example, Gibbs et al. (2018) inferred based on the sensory class of the target behavior that auditory stimuli would be matched to the reinforcers produced by vocal stereotypy. However, they also conducted a multiple stimulus without replacement PA and CSA to narrow the broad class of auditory stimuli to the specific stimulus used in their experiment (music). Given the extensive evidence that stimuli selected using PAs are likely to serve as reinforcers (e.g., Lee et al., 2010) and that stimuli effective in reducing challenging behavior in a CSA are also effective in extended assessments (Haddock & Hagopian, 2020), it seems reasonable that supplementing the matched stimulus selection method with a PA or CSA may increase success of the procedure. In fact, in a systematic review focused on self-injury, CSAs were better predictors of intervention success than preference assessments (Rooker et al., 2018).

Matched Stimulus Implementation

Just like any other stimulus, matched stimuli can be delivered on a variety of schedules such as following the omission of challenging behavior (e.g., Lanovaz & Argumedes, 2010), noncontingently (e.g., Llinas et al., 2022), or the individual could have free access to the stimulus (e.g., Love et al., 2012). In a systematic review, Rooker et al. (2018) found that NCR was the most frequently published intervention for automatically reinforced self-injurious behavior and the most effective. Although these findings may not hold when stimuli are matched (rather than arbitrary), it does provide evidence that differences in reinforcer delivery methods (e.g., DRO, NCR) can differentially affect outcomes.

Behaviors and Participants

Some variations in matched stimulus research relate directly to the target behavior or participant characteristics. For example, the topographies of behaviors targeted for reduction tend to vary across and within studies (e.g., vocal stereotypy to nail biting; Lanovaz & Sladeczek, 2011; Zawoyski et al., 2014). Considered conceptually, the topography of behavior is less likely to affect behavior reduction than the function of the behavior (Cooper et al., 2020, pp. 630). For example, Campbell (2003) conducted a qualitative synthesis of 181 individuals with autism spectrum disorder and found that the topography of the target behavior did not moderate the effects of intervention. However, this topography information may still be important for practitioners interested in using matched stimulation with their clients, who may feel more comfortable replicating a study with a dependent variable topography similar to their client's.

In addition, the method researchers use to determine that the target behavior is maintained by automatic reinforcement could differentially affect outcomes. When researchers use an assessment that does not have perfect alignment with an experimental analysis (i.e., the gold standard; Oliver et al., 2015), the confidence in their determination of function is greatly reduced. At present, the literature suggests mixed findings on this topic. For example, Campbell (2003) found that the use of an experimental functional analysis moderated intervention effects. However, in a recent study, Call et al. (2024) used functional communication training with participants whose functional behavior assessment (FBA) did and did not include an experimental analysis. They found no difference in intervention outcome based on the FBA type. We cannot be sure if these findings would hold for automatically reinforced behavior for which interventions based on matched stimulation are used.

Likewise, little is known about the individuals with whom matched stimuli have been successfully used. Wolery and Ezell (1993) note that in single-case literature, status variables (i.e., variables that do not change) such as age and disability may be less likely to affect intervention outcomes than functional variables (i.e., variables that explain current baseline performance). However, like behavioral topography, aggregating this information across matched stimulus studies would still allow for identification of the bounds of this literature (e.g., with whom have matched stimuli most often been evaluated).

Purpose

To date, no researchers have synthesized the literature on matched stimuli. As mentioned previously, some reviews did consider whether a stimulus was matched (e.g., Haddock & Hagopian, 2020), but their search criteria (i.e., must use a CSA) were unlikely to capture all matched stimulation studies. We believe a systematic review is warranted for several reasons. Aggregating studies that evaluate matched stimuli would allow practitioners interested in using matched stimuli with a client to easily identify a study similar to their context (e.g., behavioral topography, functional characteristics). In addition, this synthesis would allow us to identify potential future research directions based on identified gaps in the literature.

A meta-analysis may also be warranted to identify the overall efficacy of interventions based on matched stimulation. Haddock and Hagopian (2020) classified an 80% or greater reduction in challenging behavior as intervention success. This success metric is sometimes used in preliminary evaluations of behavioral literature (e.g., Gover et al., 2019; Rooker et al., 2013). However, this metric may ignore some practically important differences in effects; for example, when applied to a 10% increase in challenging behavior and a 70% decrease in challenging behavior, both would be scored as an "unsuccessful intervention." Last, comparing successful and unsuccessful published cases based on percent reduction lacks the statistical development necessary to serve as an effect size (Pustejovsky & Ferron, 2017).

Thus, we conducted this systematic review, meta-analysis, and moderator analyses to answer the following research questions (RQs):

RQ1: How are matched stimuli selected and used to treat automatically reinforced challenging behavior and for what types of participants and challenging behaviors?
RQ2: What is the magnitude of effect of interventions that used matched stimuli in the intervention of automatically reinforced challenging behavior relative to a control and relative to interventions that used unmatched stimuli?
RQ3: Do different assessment or treatment characteristics used in published studies influence the magnitude of intervention effects?

Method

Search and Screening Procedures

We conducted a search of the APA PsycInfo database in June 2023. We used the search terms "matched stimul*" AND "behavior" and searched any field (e.g., entire text), with no restrictions on year of publication. We used truncation to include all endings of stimul* (e.g., stimulus, stimuli, stimulation). We used quotes to ensure that matched and stimul* occurred together. This search yielded 46 articles. We conducted an initial screening requiring that articles met the inclusion criteria of: (a) peer-reviewed, (b) written in English or Spanish, and (c) published in an academic journal. Forty-one articles remained after this first screening. We then full-text screened the articles to ensure articles met the rest of the inclusion criteria, (d) focused on observable behavior, (e) included human participants, (f) involved a behavior that was maintained by automatic reinforcement, (g) intervention contained at least one replacement stimulus that the authors matched to (or attempted to match to) at least one of the same sensory systems maintaining the challenging behavior, and (h) compared the matched stimulus intervention to at least one other condition using a single-case research design. We made these determinations based on author reports. This full-text screening reduced the included articles to a total of 17.

To find articles that may have relevance to the review from before a time in which the term "matched stimuli" first emerged (Piazza et al., 2000), we also conducted an ancestral search of the references of the 17 included articles. We applied the same inclusion criteria to these 241 articles. This ancestral search yielded an additional 15 articles meeting the final inclusion criteria. Overall, a total of 32 articles met inclusion criteria (Ahearn et al., 2005; Anderson & Le, 2011; Britton et al., 2002; Carroll & Kodak, 2014; Davis et al., 2013; Favell et al., 1982; Fisher et al., 1998; Gibbs et al., 2018; Gibney et al., 2020; Higbee et al., 2005; Kennedy & Souza, 1995; Lanovaz & Argumedes, 2009; Lanovaz et al., 2009; Lanovaz & Archimedes, 2010; Lanovaz & Sladeczek, 2011; Llinas et al., 2022; Love et al., 2012; Luiselli, 1994; Piazza et al., 1998; Piazza et al., 2000; Rapp et al., 2004; Rapp, 2006; Rapp, 2007; Rapp et al., 2017; Rapp et al., 2013; Rincover et al., 1979; Saylor et al., 2012; Sidener et al., 2005; Slocum et al., 2021; Sprague et al., 1997; Taylor et al., 2005; Zawoyski et al., 2014). Figure 1 depicts these procedures.

RQ1: Descriptive Coding

We coded each article for participant descriptions, target behavior characteristics, assessments used, and matched stimulus information. First, the first author read all the articles and created a coding guide with all codes and operational definitions reported below. The research team later modified some definitions for reliability (see Interrater Agreement and Extraction Fidelity section, below) and coded the articles using the revised definitions.

Note that we only included participants in the descriptive review if during at least one experimental condition researchers provided the participant with a stimulus that researchers stated was matched (or intended to match) the sensory class (e.g., auditory) or properties (e.g., wetness) hypothesized to maintain the challenging behavior. Participants who did not receive an intervention using matched stimulation were not included in the descriptive coding. Raw data for this, and all subsequent data, are available through the online Supplementary Information.

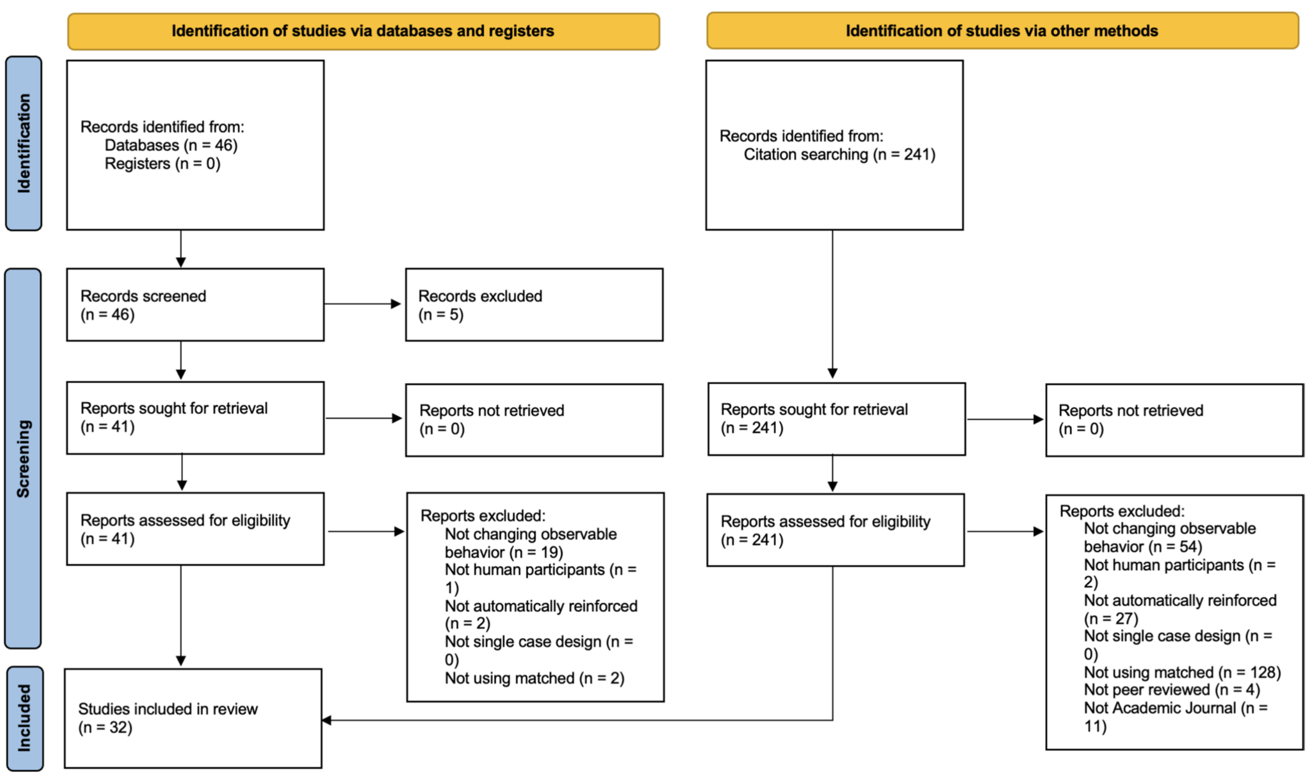

Fig. 1 PRISMA Flowchart Depicting Search and Inclusion Processes. *Note.* PRISMA = Preferred Reporting Items for Systematic reviews and Meta-Analyses. Created using Haddaway et al. (2022)

Participant Descriptions

We coded participant characteristics including gender, age, and disability or diagnosis. For gender, reviewers coded "male," "female," or "other" based on the information presented in the article (name, pronouns, explicit statements). For age, we coded participants as either 5 years old or younger, 6–12 years old, 13–17 years old, or 18–21 years old, or 22 and older. We then coded if the article reported the participant having attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder or pervasive developmental delay not otherwise specified (ASD/PDD-NOS), vision impairment, cerebral palsy, or intellectual disability. We made multiple selections for participants with multiple disabilities and could select "other" if the participants had disabilities not included in our list.

Target Behavior Characteristics

For the behavior targeted for reduction, we coded response class, and method for determination of automatic function. We categorized the target behavior from the options of dangerous climbing (any topography of behavior that results in the participant elevated from the floor, ground, or surface), self-injurious behavior (any topography of behavior that results in or has a large chance to result in damage or harm [e.g., bruising] to the participant), aggression (any topography of behavior that results in damage or harm to someone other than the participant), disruption (any topography of behavior that results in broken or damaged items), motor stereotypy (acontextual repetitive behaviors involving the body), vocal stereotypy (acontextual repetitive behaviors involving the voice), pica (any topography of behavior resulting in non-edible objects inside the mouth), or other (behaviors that do not fit into the above categories). We coded how researchers determined function by coding did not report (researchers did not state their specific method for determining function but stated it was maintained by automatic reinforcement [e.g., "not reinforced by social consequences"]), indirect assessment (researchers used interviews with other people to hypothesize the function), direct assessment (researchers made a hypothesis of function following direct observation of the behavior, but did not manipulate the environment), experimental analysis (researchers directly manipulated the environment to determine the function), or multiple.

Assessment Information

We collected data on four assessment methods used to determine whether a stimulus would match the reinforcing properties produced by the target behavior. This included: (a) inferring (researchers hypothesize the maintaining sensory class [e.g., visual], the sensory properties [e.g., brightness] the behavior produces, or both, and select a stimulus hypothesized to match this class or property), (b) unmatched comparison (researchers verify their hypothesis by comparing how assumed matched and unmatched stimuli affect the behavior), (c) MO assessment (a three-component assessment that involves collecting data directly before, during, and after the delivery of noncontingent matched stimulation; Lanovaz et al., 2010), and (d) experimental sensory analysis (any experimental analysis researchers conducted to isolate the maintaining sensory class produced by the challenging behavior, properties of the stimulus, or both). We also coded for other assessments that, although they would not help determine if the stimulus is matched, provide other information that may influence a researcher's decision to select one stimulus from among multiple matched stimuli. These included the use of a PA (any assessment that ranks stimuli in terms of participant preference) and CSA (an assessment that evaluates reduction of challenging behavior with the stimulus).

Matched Stimuli Information

We collected information on the matched stimulus delivery schedule, session duration, the number of matched stimuli used, and any other intervention procedures used in the same condition as matched stimulation. We coded NCR in the form of uninterrupted noncontingent access (NCR-C; provided the matched stimulus noncontingently at the beginning of the session and the participant had continuous access to it throughout the session), differential reinforcement of other behavior (DRO; interventionist delivered the stimulus following the omission of the target behavior for some period of time), differential reinforcement of alternative behavior (DRA; delivered contingent on an alternative behavior), NCR in the form of a fixed time delivery (NCR-FT; delivered noncontingently after a fixed period of time), and NCR in the form of variable time delivery (NCR-VT; delivered noncontingently after variable periods of time). We coded session duration by entering the duration (in minutes) of the sessions with matched stimuli. Finally, we coded whether the researchers used one matched stimulus or multiple matched stimuli during the matched stimulus condition.

RQ2: Meta-Analysis of Effects

For the meta-analysis (RQ2) and moderator analysis (RQ3), we analyzed data at the design level (whereas the descriptive coding was at the participant level) and only included designs that included at least one of our two desired comparisons: matched stimulation versus control and matched stimulation versus unmatched stimulation. We defined a control condition as an adjacent condition that contained all the same control variables as the matched stimulation condition (i.e., variables the researchers held constant across conditions; Ledford & Wolery, 2013), such as the implementer, materials, response to challenging behavior, but did not contain the intervention variables (i.e., the matched stimulus). We defined an unmatched stimulation condition as an adjacent condition that contained the same control variables as the matched stimulation condition, but the intervention variable was an unmatched stimulus rather than a matched stimulus. When researchers employed multiple designs within a study (e.g., conducting an MO analysis then later using the stimuli from the MO analysis to create a classroom intervention), we used only the final (i.e., conducted last) experimental design in each manuscript for each participant. If the final experimental design for each participant did not have a comparison that met either requirement (e.g., the final design only compared matched stimulation adjacent to RIRD), we repeated this process with the prior experimental design in the study until we identified one experimental design that met this criteria or there were no more designs left in the study to code. The figure number associated with the included design for each study can be found in the online Supplementary Information.

Calculating Metrics

Using WebPlotDigitizer (Drevon et al., 2017), we extracted the raw data for meta-analytic comparisons (e.g., control vs. matched) if there were three or more potential demonstrations of effect (i.e., three manipulations of the IV at three different points in time). For example, an A-B-A design would not be included because the researcher only introduced and withdrew the intervention, constituting only two manipulations. However, an A–B–A–B design includes an extra introduction of the intervention, meeting the three manipulations requirement. If a researcher compared a control condition (A), a matched stimulus condition (B), and unmatched stimulus condition (C) in an A–B–A–B–C–B–C design, we selected the "A–B–A–B" to compare matched stimulation versus a control condition and "B–C–B–C" to compare matched stimulation versus unmatched stimulation. We did not extract maintenance, follow-up, or generalization data, but did include schedule thinning data in the one study in which it occurred during our target comparisons (Slocum et al., 2021).

We calculated log response ratio for interventions decreasing behavior effect sizes (LRRd; Pustejovsky & Ferron, 2017). We selected LRRd as our effect size because it is appropriate for behaviors measured on a ratio scale (i.e., a value of zero means absence of the behavioral construct) regardless of the measurement system (Pustejovsky, 2018). In our case, we had a unifying construct (i.e., challenging behavior) that, due to study level differences, was measured differently across studies (e.g., different operational definitions and different measurement systems). LRR as an effect size quantifies the direction (i.e., increase or decrease) and proportion of change (i.e., change relative to baseline performance) between conditions and can be used for meta-analysis of single case designs (Moeyaert et al., 2018). Thus, for LRR, a negative value represents decreases in challenging behavior and the larger the negative values represent larger proportional decrease in behavior. In addition, LRRd can be mathematically converted to the commonly used metric mean baseline percentage reduction (Haddock & Hagopian, 2020; Pustejovsky, 2018). Published literature suggests that this effect size should not be used when behavior change is expected to occur gradually (Pustejovsky, 2018), which conceptually is unlikely for matched stimuli.

One unique consideration with LRR is that the effect size is calculated for a single A–B comparison (e.g., baseline phase followed by intervention phase), thus researchers have several options for calculating phase contrasts when multiple A–B comparisons are made (Pustejovsky & Ferron, 2017; Pustejovsky, 2018). The first option is to take only the data from the first A–B contrast; this option is limited because it ignores some of the data. The second option is to aggregate all the data from A phases and all the data from B phases, then calculate LRR. This method should only be used when changes in level are the same across A–B comparisons (Pustejovsky, 2018). The last method is to calculate LRR for the first A–B, the second A–B comparison, then average the two effect sizes. This method allows researchers to draw conclusions without assuming that changes in the dependent variable are consistent across A–B comparisons. In this study, we elected to use the second method. Our visual analyses suggested that changes in behavior during A–B comparisons were consistent within studies, making the second method of selecting phase contrast strategies functionally equivalent to the third (calculating an effect size for each A–B comparison then averaging these effects). To calculate effect sizes, we used the R package *SingleCaseES* (Pustejovsky et al., 2023), and programmed the outcome scale (e.g., rate, percentage) for each study. R code for this, and all subsequent analyses, is available via the online Supplementary Information.

Overall Effect Calculation

Given that our data were structured hierarchically (i.e., cases within studies), we conducted a multilevel meta-analysis (Van den Noortgate & Onghena, 2007) to calculate an overall effect size using LRRd. We selected this model because, unlike basic meta-analysis models that average out case-level variability, a multilevel model can provide an overall average effect size while considering variation at the case and study levels (e.g., Ganz et al., 2023). We used the cluster-robust variance estimation (Chen & Pustejovksy, 2022) to control for the chance of inaccurate sampling variances due to autocorrelation (i.e., serial dependence) of single case research design data and to adjust for the small sample size (Pustejovsky, 2018),

To conduct this multilevel meta-analysis, we used the R packages *Metafor* (Pustejovsky & Ferron, 2017; Viechtbauer, 2010), and *clubsandwich* for cluster-robust variance estimation (Pustejovsky, 2022). In particular, we used the LRRd data, the squared standard error as the variance, and clustered the results by study. We also calculated 80% prediction intervals. Prediction intervals provide additional data beyond an effect size, reflecting the heterogeneity of meta-analysis results by showing the expected range of true effects in similar studies (InHout et al., 2016). An 80% prediction interval informs what range results would fall in, with 80% confidence, if one used matched stimulation as an intervention in a new sample.

RQ3: Moderator Analyses (Post-Hoc)

We conducted moderator analyses to inform the extent to which case-level characteristics explained variation in the magnitude of the effect of matched stimulus interventions on challenging behavior. A moderator is a variable that affects the strength of the relation between the independent and dependent variables (Hall & Rosenthal, 1991). Moderator analyses can inform for whom and under what conditions an intervention is effective. All variables we coded for in descriptive coding were considered as potential moderators (i.e., gender, disability, age, behavior type, FBA type, matched stimulus determination method, PA usage, CSA usage, intervention type, number of stimuli, and session duration). However, to avoid inflating the cumulative error by conducting analyses on each variable, we selected variables to test as moderators if there was (a) a theoretical reason this variable might affect the success of the intervention and (b) the variable contained a sufficient number of comparisons at each level (i.e., five or more per level).

We programmed the model with separate intercepts for each level (i.e., category) of the moderator and tested each moderator individually (Pustejovsky & Tipton, 2022). In addition, we ran linear contrasts to show the differences between levels of the moderator when the overall test was significant (Mok et al., 2023). We conducted moderator analyses using the *Metafor* and *clubsandwich* R-packages and included a small sample correction for the *p*-value (Pustejovsky, 2022).

Interrater Agreement and Extraction Fidelity

We took several steps to ensure the accuracy of the data included in this review. A secondary evaluator (the second author) was trained on article screening and descriptive coding. For each, she was given one practice article, informed how to score each component, and given feedback on that article after she coded it. The secondary evaluator screened a random 20 articles from the primary search (43.48%) and 78 articles from the ancestral search (31.54%) and obtained 100% and 96.05% IRR, respectively, with the first author (using a point-by-point method; Ledford et al., 2018). For the descriptive coding (RQ1), the secondary evaluator coded 10 randomly selected articles (31.25%) and obtained 82.72% agreement. After this first coding, we elected to add another variable we thought might influence outcomes (matched stimulus delivery schedule), removed a variable that had not been validated for any behaviors except self-injury (automatic subtype), and adjusted any definitions that the primary or secondary evaluator reported were unclear. The final definitions we used are the ones reported in this method section. The primary evaluator and secondary evaluator independently recoded the articles and obtained 91.74% agreement. To resolve any disagreements, the two evaluators discussed and reached consensus.

To ensure fidelity in data extraction, the extractor saved a picture of their desktop showing the WebPlotDigitizer completed page (e.g., graph, selected data points, selected x-values). The secondary evaluator viewed these screenshots and scored whether the extractor: (a) accurately selected a clearly labeled low point and high point on the Y axis (b) accurately input the y-axis values (c) accurately selected a clearly labeled low point and high point on the x-axis, (d) accurately input the x-axis values into WebPlotDigitizer (e) accurately selected the data points representing the comparison condition, (f) accurately selected the data points representing the target condition, and (g) accurately select only conditions that are adjacent. The extractor completed all steps with 100% fidelity.

Results

RQ1: Descriptive Coding

This review included 32 studies, which contained 71 participants. Descriptive data are summarized in Tables 1, 2, and 3, and associated raw data are available in the online Supplementary Information. Of these 71 participants, 51 (71.83%)

were male and 20 (28.16%) were female. The participants had various disabilities including 26 (36.62%) with intellectual disability, 51 (71.83%) with ASD/PDD-NOS, 5 (7.04%) with vision impairment, 3 (4.23%) with ADHD, 4 (5.63%) with no reported diagnoses, and 17 (23.94%) with multiple disabilities. Fourteen participants (19.72%) were between the ages of 0 and 5, 35 (49.30%) between the ages of 6 and 12, 12 (16.90%) between 13 and 17, 4 (5.63%) between 18 and 21, and 3 (4.23%) were 22 years old or older. For three participants (4.23%) the authors did not provide the age.

Researchers targeted a variety of behaviors including vocal stereotypy for 31 participants (43.66%), motor stereotypy for 20 (28.17%), self-injurious behavior for 7 (9.86%), pica for 8 (11.27%), disruption for 2 (2.87%), dangerous climbing for 1 (1.41%), and aggression for 0 participants. We had five behaviors coded as other (7.04%; 1 hand mouthing, 3 nail biting, and 1 saliva play). For 3 participants, researchers targeted two different behaviors. Researchers determined behavioral function using indirect assessment for 3 participants (4.22%), direct assessment for 13 participants (18.31%), experimental analysis for 38 participants (53.52%), and a combination of methods for 3 participants (4.23%). Last, we found that for 20 participants (28.17%), researchers did not report how they came to the automatic reinforcement conclusion.

In the selection or evaluation of the matched stimuli, researchers inferred with 28 participants (39.44%), used an unmatched comparison with 9 participants (12.68%), used MO assessment with 14 participants (19.72%), and used experimental sensory analysis with 23 participants (32.39%). Researchers used multiple methods of selecting matched stimuli with 3 participants (4.26%). Researchers used a PA with 41 participants (57.75%) and a CSA with 13 (18.31%).

For intervention schedules, researchers used NCR-C with 67 participants (94.36%), DRO with 6 participants (8.45%), DRA with 2 participants (2.82%), NCR-FT with 4 participants (5.63%), and NCR-VT with 2 participants (2.82%). Ten (14.08%) of these participants received more than one of these across experimental conditions. The duration of sessions was 5 min for 21 participants (29.56%), 8 min for 2 (2.82%) participants, 10 min for 34 (47.89%), 15 min with 7 participants (9.86%), 17 min with 1 participant (1.41%), 20 min with 2 participants (2.82%), and 30 min with 4 participants (42.25%). Researchers used one stimulus with 40 participants (56.34%), multiple stimuli with 32 (45.07%), and one participant (1.41%) received both one and multiple matched stimuli across experimental conditions.

Table 1 Participant Descriptions

	$n = 71$	Percent
Gender		
Male	51	71.83
Female	20	28.16
Age		
0–5	14	19.72
6–12	35	49.30
13–17	12	16.90
18–21	4	5.63
22+	3	4.23
Did not state	3	4.23
Diagnosis [a]		
ID	26	36.62
ASD/PDD-NOS	51	71.83
Vision Impairment	5	7.04
ADHD	3	4.23
None	4	5.63

Note. ID = intellectual disability; ASD = autism spectrum disorder; PDD-NOS = pervasive developmental disorder, not otherwise specified; ADHD = attention-deficit/hyperactivity disorder

[a]16 participants (22.54%) had two disabilities and one participant (1.41%) had three disabilities, thus these data exceed 100%

Table 2 Challenging Behavior Characteristics

	n	Percent
Challenging Behaviors Targeted [a]		
Vocal stereotypy	31	43.66
Motor stereotypy	20	28.17
Self-injury	7	9.86
Pica	8	11.27
Disruption	2	2.87
Dangerous climbing	1	1.41
Aggression	0	0
Other	5	7.04
Functional Behavior Assessment Method [b]		
Did not report	20	28.17
Indirect assessment	3	4.23
Direct assessment	13	18.31
Experimental analysis	38	53.52
Combination	3	4.23

Note. [a] Three participants (4.23%) had two behaviors targeted, thus these data exceed 100%

[b] Three participants (4.23%) used two methods in combination to determine function, thus these data exceed 100%

RQ3: Meta-Analysis of Effects

Although the descriptive review contained 71 participants from 32 studies, only 49 cases across 20 studies met criteria for comparing the matched stimulus condition to a control condition, and only 14 cases across seven studies met criteria for comparing the matched stimulus condition to an unmatched stimulus condition. Results of the multilevel meta-analysis evaluating the effect of matched stimuli

Table 3 Matched Stimuli Information

	$n = 71$	Percent
Determination of "Matched" [a]		
Inferring	28	39.44
Inferring then compare to unmatched	9	12.68
Inferring then evaluate MOs surrounding usage	14	19.72
Experimental sensory analysis	23	32.39
Assessments Used[b]		
Preference	41	57.75
Competing stimuli	13	18.31
None of the above	19	26.76
Schedules of Intervention[c]		
NCR-C	67	94.36
DRO	6	8.45
DRA	2	2.82
NCR-FT	4	5.63
NCT-VT	2	2.82
Duration of Session		
5 min	21	29.56
8 min	2	2.82
10 min	34	47.89
15 min	7	9.86
17 min	1	1.41
20 min	2	2.82
30 min	4	5.63
Number of Stimuli Used[d]		
One	40	56.34
Multiple	32	45.07

Note. MOs = motivating operations; NCR-C = noncontingent reinforcement-continuous; DRO = differential reinforcement of other behavior; DRA = Differential reinforcement of alternative behavior; NCR-FT = noncontingent reinforcement-fixed time; NCR-VT = noncontingent reinforcement-variable time

[a] Three participants (4.23%) used multiple methods, thus these data exceed 100%

[b] Two participants (2.82%) used both preference and CSA, thus these data exceed 100%

[c] Ten participants (14.08%) used multiple schedules, thus these data exceed 100%

[d] One participant (1.41%) used both, thus these data exceed 100%

relative to control (Fig. 2) show an average effect size of LRRd = −1.99 (95% CI [−2.62, −1.36]), which was statistically significant ($p < .0001$). The results of our prediction interval suggest, with 80% confidence, that a new participant in a new sample would demonstrate between a 98.38% decrease and a 14.67% increase in challenging behavior. Results of the meta-analysis evaluating matched stimuli relative to unmatched stimuli (Fig. 3) showed an average effect size of LRRd = −1.64 (95% CI [−2.79, −0.48]), which was statistically significant ($p = .0141$). The results of the prediction interval suggest that, with 80% confidence, a new participant would demonstrate between a 97.35% decrease and a 43.15% increase in challenging behavior. Both prediction intervals are wide due to large case-level variability.

RQ4: Moderator Analysis

Table 4 contains a summary of the process for identifying possible moderator variables. During a meeting with the first and third author, we excluded gender, disability, behavior type, and session duration from the moderator analysis because we could not identify a conceptual reason that the efficacy of matched stimuli would differ according to these factors. We excluded age from moderator analyses due to limited variability in age; the sample only included two cases with participants in the 18–21 range and one case in the 22+ range. We excluded the matched stimulus determination method from moderator analyses because the sample only included three cases in which researchers used an experimental method. Last, for intervention type, only two cases delivered stimuli using a DRO procedure and only one case used an NCR-FT delivery. Thus, neither age, matched stimulus determination method, nor intervention type were included in moderator analyses because they did not have at least five cases per level (i.e., category).

We did not find a statistically significant effect for researcher's use of a CSA ($F(1, 3.86) = 1.1223$, $p = .3511$) or a PA ($F(1, 10.2) = 3.7724$, $p = .0802$) as moderator variables. A larger sample size might allow researchers to detect smaller changes in performance when these assessments are used. However, we found statistically significant effects for the number of matched stimuli used ($F(1, 13.48) = 5.9142$, $p = .0296$). When including the number of matched stimuli as a moderator, the estimated average effect size was LRRd = −1.4183 ($p = .0040$; 75.9% decrease in challenging behavior) for one matched stimulus and LRRd = −1.2433 ($p = .0296$; 71.2% decrease in challenging behavior) for multiple matched stimuli. These results indicate a 4.7% greater decrease in challenging behavior when using one matched stimulus compared to multiple matched stimuli.

With FBA type included as a moderator, the overall test showed mean LRRd differed significantly across FBA categories ($F(3, 6.9) = 11.0987$, $p = 0.0049$). The estimated effect size was LRRd = −1.765 ($p = .0010$, 95% CI [−2.5984, −0.9330]) for experimental assessment, LRRd = −2.4841 ($p = .0119$, 95% CI [−4.1338, −0.8343]) for direct assessment, and LRRd = −1.8520 ($p = .1247$, 95% CI [−4.6826, 0.9786]) for studies that did not report this information. Both experimental assessment and direct assessment had statistically significant negative effect sizes, indicating these conditions were associated with a greater decrease in challenging behavior when compared to a control condition. The "not reported" category

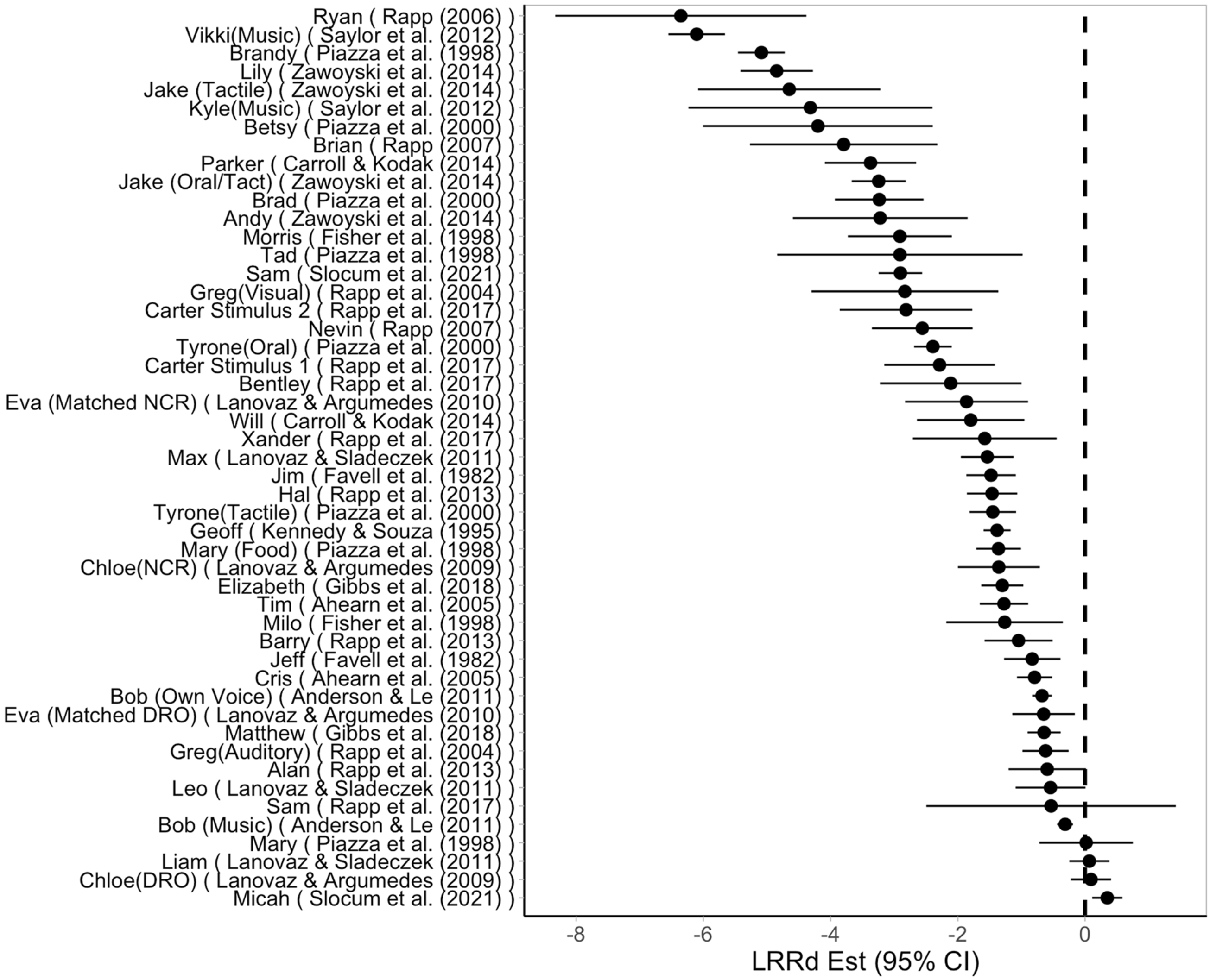

Fig. 2 Forest Plot of Cases Comparing Matched Stimulation versus Control. *Note*. LRRd = log response ratio for interventions decreasing behavior. LRRd scores at zero depict studies with no effect, LRRd values to the right of zero (positive values) suggest an increase in interfering behavior compared to baseline, those to the left of zero (negative values) suggest decrease in behavior compared to baseline. The further to the left the data point (the more negative), the larger the magnitude of the decrease in behavior. Horizontal bars represent 95% confidence intervals

indicated a large but statistically nonsignificant average reduction in challenging behavior. The wide confidence interval suggests imprecise estimation, which likely reflects variability in this subgroup or the small number of studies. Thus, although effects in this category are directionally similar to the other two categories, conclusions about magnitude should be made cautiously. These data taken together suggest that verifying the automatic function with either experimental or direct assessment may be an important variable for ensuring the effectiveness of matched stimuli. Because the omnibus *F* test was significant, and we had multiple levels of the moderator, we conducted linear pairwise comparisons to test whether group means differed significantly. Pairwise comparisons using cluster-robust variance estimation showed no statistically significant differences between FBA types, though the pattern of means suggested somewhat larger decreases in challenging behavior following interventions informed by descriptive FBAs compared to other types of FBAs.

Discussion

When challenging behavior is maintained by automatic reinforcement, intervention can become complicated because the implementer does not have control of the reinforcer delivery (Vaughan & Michael, 1982; Vollmer, 1994). Because interventionists are unable to isolate the

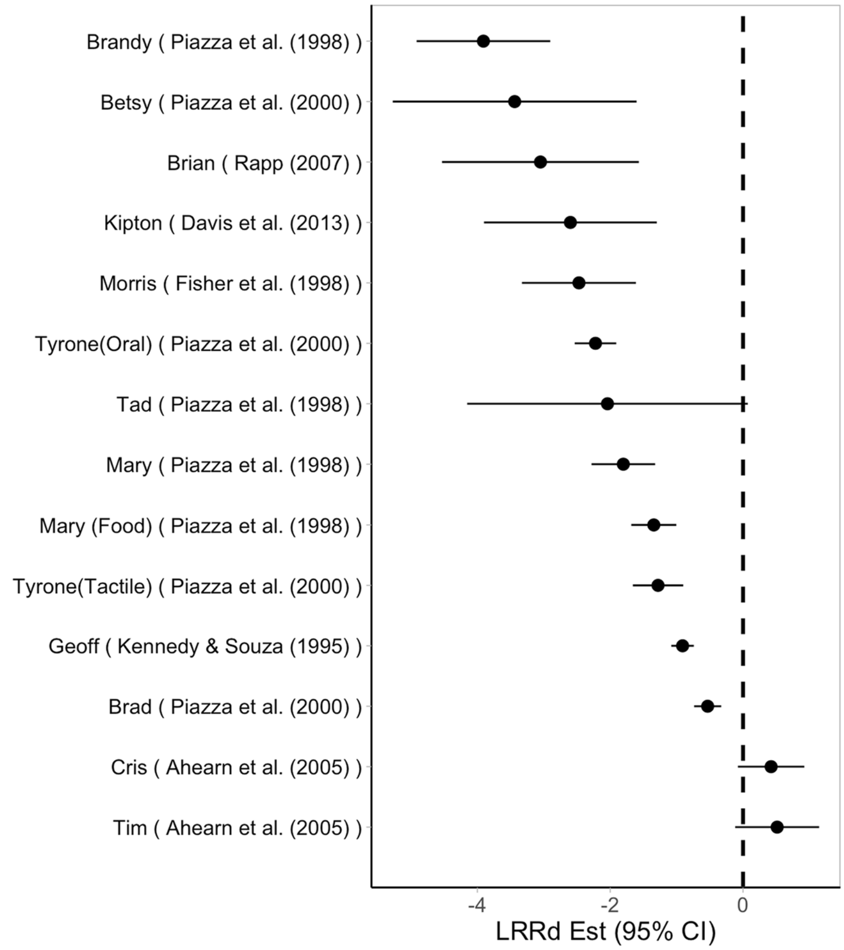

Fig. 3 Forest Plot of Cases Comparing Matched to Unmatched Stimuli. *Note*. LRRd = log response ratio for interventions decreasing behavior. LRRd scores at zero depict studies with no effect, LRRd values to the right of zero (positive values) suggest an increase in behavior compared to unmatched, those to the left of zero (negative values) suggest decrease in behavior compared to unmatched. The further to the left the data point (the more negative), the larger magnitude the decrease in behavior. Horizontal bars represent 95% confidence intervals

specific reinforcers produced by the challenging behavior, behavior analysts have often focused on identifying stimuli that compete with the behavior and presenting them noncontingently. However, there may be a conceptual difference when stimuli that provide the same source of reinforcement as the challenging behavior are used (i.e., matched stimuli). If the stimulus is truly matched in this way, engaging with the stimulus may functionally serve as a replacement behavior. Recent research suggests that this difference (i.e., used of matched or unmatched stimuli) made little difference on challenging behavior reduction (Haddock & Hagopian, 2020); however, limitations of this research suggest a need for further inquiry.

We conducted a systematic literature review of experimental articles that used matched stimulation to reduce challenging behavior maintained by automatic reinforcement. Our search methods identified 32 published articles using matched stimulation with a total of 71 participants. We coded these published cases of matched stimulation to: (a) determine specific participant and intervention variables present in the literature, (b) calculate the overall average effect size of matched stimulus interventions in these studies, and (c) determine if any intervention variables moderate these effects.

Average Effect Size

The results of our multilevel meta-analysis found the average effects of interventions using matched stimuli to be statistically significant when compared to a control condition. In other words, matched stimuli resulted in reductions in challenging behavior relative to a control condition that are unlikely to be due to chance. The average effect size was LRRd = −1.99, which corresponds to an 86.47% decrease in challenging behavior on average across all cases. In addition, when compared to interventions using unmatched stimuli, we found the average effect of matched stimuli to be LRRd = −1.64. Thus, on average, interventions using matched stimuli resulted in an 80.60% decrease in challenging behavior relative to interventions using unmatched stimuli. These findings support the use of interventions with matched stimuli to treat automatically reinforced challenging behavior. In addition, our data suggest that practitioners should use matched stimuli, rather than arbitrary (possibly unmatched) stimuli when possible.

However, there were four cases in our data set in which matched stimuli were ineffective or countertherapeutic (i.e., LRRd ≥ 0) relative to a control condition, suggesting that

Table 4 Process for Selecting Potential Moderators to Test

Variables Coded	Conceptual Reason to Test?	Sufficient Cases in Each Level to Test?	Included as Potential Moderator?
Gender	No	Yes; $n = 11$	No
Disability	No	No; $n = 4$	No
Age	Yes; potentially a longer learning history with the behavior	No; $n = 1$	No
Behavior	No	No; $n = 1$	No
FBA Method	Yes; see intro	Yes; $n = 8$	Yes
Matched determination method	Yes; see intro	No; $n = 3$	No
PA	Yes; see intro	Yes; $n = 17$	Yes
CSA	Yes; see intro	Yes; $n = 11$	Yes
Intervention	Yes; see intro	No; $n = 1$	Yes
Number of stimuli	Yes; see intro	Yes; $n = 23$	Yes
Session length	No	Yes; Continuous variable	No

Note. FBA = functional behavior assessment; PA = preference assessment; CSA = competing stimulus assessment

practitioners should continue to take an idiographic approach with their individual clients; no intervention will work for all individuals in all contexts. This finding should be contextualized with the fact that most of these studies included only participants with developmental disabilities (94.37%). In addition, these results should be taken into consideration with our other moderator findings, limitations, and gaps in the current research.

Matched Stimulus Selection

Our descriptive review revealed that for a large portion of participants ($n = 28$; 39.44%), researchers simply inferred that the stimulus was matched. In these cases, the researcher hypothesized what sense modality (e.g., tactile) and stimulus property (e.g., hard) might be maintaining the challenging behavior and chose a stimulus that they inferred may be matched to that modality or property. The lack of systematic methods for selecting matched stimuli may limit replicability, which is an essential part of science (Tincani & Travers, 2019).

In fact, the differences in how matched stimuli are selected may be critical in explaining and enhancing the effectiveness of matched stimuli, given that each method of identifying matched stimuli has its own strengths and limitations. For example, the MO assessment for selecting a matched stimuli evaluates whether the presence of the matched stimulus acts as an abolishing operation for the target challenging behavior. This method assumes that one session of access to a functionally matched stimulus will reduce the establishing operation for challenging behavior to a degree that can be measured. In reality, automatically reinforced behavior may not satiate in every case (e.g., Harlow 1950, 1956). In conceptual terms, we would expect experimental analyses to be more likely to identify stimuli that are truly, functionally, matched. Unfortunately, even though our descriptive coding found that researchers used an experimental analysis to select the matched stimulus with 23 participants (32.39%), only 3 of these cases were included in the meta-analysis, due to the evaluations not including the comparisons we were interested in (e.g., matched stimulation vs. unmatched stimulation or matched stimulation vs. baseline) or not having three potential demonstrations of effect (e.g., A–B design). Due to a limited number of studies on each assessment type, we were unable to use this variable in our moderator analysis. However, our data suggest that use of matched stimuli, on average, was effective regardless of the method of identifying these stimuli. Future research could investigate this question more directly, for example by comparing stimuli selected with each method (e.g., infer, compare to hypothesized unmatched, MO, and experimental assessment). In addition, we hypothesized that the effects of matched stimuli would be greater when supplemented with other validated assessments such as a PA or CSA (Rooker et al., 2018). However, our moderator analysis revealed no statistical significance for PA use nor CSA use. More research is needed to understand how the assessment methods used to select stimuli (based on preference, response competition, or matched status) affect the success of NCR.

Behaviors and Participants

For 20 participants (28.17%), authors did not report how they concluded that the behavior was maintained by automatic reinforcement. Given that matched stimulation is conceptually linked to automatic reinforcement, if authors make a false determination of function, it may affect outcomes. Our moderator analysis verified this concern; we found that FBA type moderated outcomes. The data pattern suggests that there may be a larger decrease in challenging behavior when researcher use experimental or descriptive assessments compared to if the assessment is not reported. However, because pairwise comparisons were not significant (confidence intervals included zero), we need a larger sample to increase confidence in these findings. In general, these findings suggest FBA may be important to verifying the automatic reinforcement function systematically. We suggest future researchers continue to compare outcomes when using different FBA methodologies to extend these findings.

Matched Stimulus Implementation Approaches

The results of our descriptive coding also revealed a few interesting patterns about the types of interventions used with matched stimuli. First, although there are many different schedules of reinforcement researchers use to deliver matched stimuli (e.g., DRO, NCR-FT, NCR-C), 94.36% of participants received NCR-C. Thus, practitioners should note NCR-C as the most researched schedule for delivering matched stimuli and future researchers should evaluate the effects of matched stimuli when delivered via other schedules. Unfortunately, we did not have sufficient examples to test differences in reinforcement schedule delivery within our moderator analysis.

The moderator analysis revealed a statistically larger effect for comparisons that used one matched stimulus over those that used multiple matched stimuli. These findings are similar to Clay et al. (2018), who found that providing noncontingent access to one stimulus (identified via CSA) was either equally effective as or more effective than using multiple stimuli. We initially hypothesized that multiple stimuli would result in stronger effects as there might be a larger chance that at least one of the stimuli is functionally matched. One potential interpretation for these findings is that when multiple stimuli are used, they may not all be "equally" matched. In general, the basic literature suggests that one sense modality (e.g., visual stimuli) may not be discrete enough for a functional classification, as there are published examples in which stimuli within the same sense modality (e.g., visual or stimuli) may have differences in reinforcer potency (e.g., Barnes & Baron, 1961; Baron & Kish, 1962). Given the exploratory nature of this analysis and that the results were not what we expected, we encourage future researchers to continue to directly evaluate this effect.

Limitations

One limitation of this review is that when researchers conducted an assessment to select matched stimuli, they did not always conduct a subsequent evaluation of the effectiveness of the stimulus selected from the assessment. For example, Rapp (2006) conducted an MO assessment but did not select a stimulus from this assessment nor conduct a subsequent analysis. Although we could use these data in our meta-analysis, we could not accurately characterize this comparison as *informed by* an MO assessment, because the data included in our meta-analysis were *from* the MO assessment. Thus, the stimulus used during this assessment was actually selected by inference. Overall, this type of issue limited the conclusions we could draw from this sample about the role of matched stimulus assessment type. Thus, although these assessments are present in the literature, there were insufficient evaluations of matched stimuli selected from these assessments for us to include this as a variable in the moderator analysis. Future researchers could consider directly comparing the outcomes of matched stimuli selected using each of the assessment methodologies.

Another limitation of this review is that we used relatively narrow search terms to identify articles (e.g., "matched stimul*" and behavior). It is possible that using broader search terms commonly associated with matched stimulus interventions (e.g., environmental enrichment, noncontingent reinforcement) would have resulted in identifying more articles. We used an ancestral search to identify relevant articles that may not have used terms like matched stimuli or matched stimulation (e.g., Favell et al., 1982). We found 46.88% ($n = 15$) of our studies using this ancestral search, suggesting that our search terms did not capture all relevant articles. Likewise, excluding theses and dissertations (i.e., grey literature) may have biased our findings in a positive direction (Tincani & Travers, 2019). Future researchers may consider expanding their search criteria to include other keywords and databases beyond APA PsycINFO and including grey literature.

Conclusions

Overall, our data suggest interventions using matched stimuli are an effective method for reducing challenging behavior maintained by automatic reinforcement (when compared to a control condition or a condition using unmatched stimuli). In addition, because we found matched stimuli to be significantly more effective than unmatched stimuli, we recommend practitioners who are considering using competing stimulation to also consider matching the function of the stimulus such that engagement with the stimulus may provide the same stimulation as, and thus replace the consequences of, the challenging behavior. Based on our moderator analysis, practitioners are more likely to have success with matched stimulation if they use one stimulus at a time and verify the automatic reinforcement function of the challenging behavior using either direct or experimental

FBA methods. More research is needed to further refine and evaluate the methods for identifying and selecting matched stimuli (e.g., inferring, MO assessment, unmatched comparison, experimental sensory analysis) and how different schedules of matched stimulus delivery (e.g., NCR-C, NCR-FT, DRO) may influence outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40617-025-01140-2.

Author Contributions All authors contributed to the conceptualization, implementation, and/or writing of the article.

Funding This work was partially supported by the National Institute of Mental Health award 1R56MH129489-01A1.

Data Availability Supplementary Materials 1 contains all raw data and Supplementary Materials 2 contains all R code and output. These supplemental documents are available with this article on the publisher website.

Declarations

Conflicts of Interest The authors have no competing interests to declare.

Ethical Approval This work did not utilize human participants and thus does not require ethics committee approval nor an informed consent process.

References

* = Study included in the systematic review.

*Ahearn, W. H., Clark, K. M., DeBar, R., & Florentino, C. (2005). On the role of preference and response competition. *Journal of Applied Behavior Analysis, 38*, 247–250. https://doi.org/10.1901/jaba.2005.36-04

*Anderson, J., & Le, D. D. (2011). Abatement of intractable vocal stereotypy using an overcorrection procedure. *Behavioral Interventions, 26*(2), 134–146. https://doi.org/10.1002/bin.326

Barnes, G. W., & Baron, A. (1961). Stimulus complexity and sensory reinforcement. *Journal of Comparative and Physiological Psychology, 54*(4), 466–469. https://doi.org/10.1037/h0046708

Baron, A., & Kish, G. B. (1962). Low-intensity auditory and visual stimuli as reinforcers for the mouse. *Journal of Comparative and Physiological Psychology, 55*(6), 1011–1013. https://doi.org/10.1037/h0044286

Berg, W. K., Wacker, D. P., Ringdahl, J. E., Stricker, J., Vinquist, K., Salil Kumar Dutt, A., Dolezal, D., Luke, J., Kemmerer, L., & Mews, J. (2016). An integrated model for guiding the selection of treatment components for problem behavior maintained by automatic reinforcement. *Journal of Applied Behavior Analysis, 49*(3), 617–638. https://doi.org/10.1002/jaba.303

*Britton, L. N., Carr, J. E., Landaburu, H. J., & Romick, K. S. (2002). The efficacy of noncontingent reinforcement as treatment for automatically reinforced stereotypy. *Behavioral Interventions, 17*, 93–103. https://doi.org/10.1002/bin.110

Call, N. A., Bernstein, A. M., O'Brien, M. J., Schieltz, K. M., Tsami, L., Lerman, D. C., Berg, W. K., Lindgren, S. D., Connelly, M. A., & Wacker, D. P. (2024). A comparative effectiveness trial of functional behavioral assessment methods. *Journal of Applied Behavior Analysis, 57*(1), 166–183. https://doi.org/10.1002/jaba.1045

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing challenging behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24*(2), 120–138. https://doi.org/10.1016/s0891-4222(03)00014-3

Carr, J. E., Coriaty, S., Wilder, D. A., Gaunt, B. T., Dozier, C. L., Britton, L. N., Avina, C., & Reed, C. L. (2000). A review of "noncontingent" reinforcement as treatment for the aberrant behavior of individuals with developmental disabilities. *Research in Developmental Disabilities, 21*(5), 377–391. https://doi.org/10.1016/s0891-4222(00)00050-0

*Carroll, R. A., & Kodak, T. (2014). An evaluation of interrupted and uninterrupted measurement of vocal stereotypy on perceived treatment outcomes. *Journal of Applied Behavior Analysis, 47*(2), 1–13. https://doi.org/10.1002/jaba.118

Chen, M., & Pustejovsky, J. E. (2022). Multilevel meta-analysis of single-case experimental designs using robust variance estimation. *Psychological Methods.*https://doi.org/10.1037/met0000510

Clay, C. J., Clohisy, A. M., Ball, A. M., Haider, A. F., Schmitz, B. A., & Kahng, S. (2018). Further evaluation of presentation format of competing stimuli for treatment of automatically maintained challenging behavior. *Behavior Modification, 42*(3), 382–397. https://doi.org/10.1177/0145445517740322

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied behavior analysis* (3rd ed). Pearson Education.

*Davis, T. N., Dacus, S., Strickland, E., Machalicek, W., & Coviello, L. (2013). Reduction of automatically maintained self-injurious behavior utilizing noncontingent matched stimuli. *Developmental Neurorehabilitation, 16*(3), 166–171. https://doi.org/10.3109/17518423.2013.766819

Dessemontet, R., Geyer, M., Linder, A.-L., Atzemian, M., Martinet, C., Meuli, N., Audrin, C., & de Chambrier, A.-F. (2024). Effects of shared text reading for students with intellectual disability: A meta-analytical review of instructional strategies. *Educational Research Review, 44*, 100615. https://doi.org/10.1016/j.edurev.2024.100615

Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behavior Modification, 41*(2), 323–339. https://doi.org/10.1177/0145445516673998

*Favell, J. E., McGimsey, J. F., & Schell, R. M. (1982). Treatment of self-injury by providing alternate sensory activities. *Analysis and Intervention in Developmental Disabilities, 2*(1), 83–104. https://doi.org/10.1016/0270-4684(82)90007-6

*Fisher, W. W., Lindauer, S. E., Alterson, C. J., & Thompson, R. H. (1998). Assessment and treatment of destructive behavior maintained by stereotypic object manipulation. *Journal of Applied Behavior Analysis, 31*, 513–527.

Ganz, J. B., Pustejovsky, J. E., Reichle, J., Vannest, K. J., Foster, M., Pierson, L. M., Wattanawongwan, S., Bernal, A. J., Chen, M., Haas, A. N., Liao, C.-Y., Sallese, M. R., Skov, R., & Smith, S. D. (2023). Participant characteristics predicting communication outcomes in AAC implementation for individuals with ASD and IDD: A systematic review and meta-analysis. *Augmentative and Alternative Communication, 39*(1), 7–22. https://doi.org/10.1080/07434618.2022.2116355

*Gibbs, A. R., Tullis, C. A., Thomas, R., & Elkins, B. (2018). The effects of noncontingent music and response interruption and redirection on vocal stereotypy. *Journal of Applied Behavior Analysis, 51*(4), 899–914. https://doi.org/10.1002/jaba.485

*Gibney, C., Phillips, K. J., Arnold, S. A., & Taylor, S. A. (2020). An evaluation of the effects of response interruption redirection and matched stimulation on vocal stereotypy. *Behavioral Interventions, 35*(1), 114–130. https://doi.org/10.1002/bin.1700

Gover, H. C., Fahmie, T. A., & McKeown, C. A. (2019). A review of environmental enrichment as treatment for problem behavior maintained by automatic reinforcement. *Journal of Applied Behavior Analysis, 52*(1), 299–314. https://doi.org/10.1002/jaba.508

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews, 18*, Article e1230. https://doi.org/10.1002/cl2.1230

Haddock, J. N., & Hagopian, L. P. (2020). Competing stimulus assessments: A systematic review. *Journal of Applied Behavior Analysis, 53*(4), 1982–2001. https://doi.org/10.1002/jaba.754

Hall, J. A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: Issues and methods. *Communication Monographs, 58*(4), 437–448. https://doi.org/10.1080/03637759109376240

Harlow, H. F. (1950). Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. Journalof Comparative and Physiological Psychology, 43(4), 289. https://doi.org/10.1037/h0058114

Harlow, H. F., Blazek, N. C., & McClearn, G. E. (1956). Manipulatory motivation in the infant rhesus monkey. Journal of Comparative and Physiological Psychology, 49(5), 444. https://doi.org/10.1037/h0047817

*Higbee, T. S., Chang, S., & Endicott (2005). Noncontingent access to preferred sensory stimuli as a treatment for automatically reinforced stereotypy. *Behavioral Interventions*, *20*, 177–184. https://doi.org/10.1002/bin.190

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open, 6*(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis, 27*(2), 197–209. https://doi.org/10.1901/jaba.1994.27-197

*Kennedy, C. H., & Souza, G. (1995). Functional analysis and treatment of eye poking. *Journal of Applied Behavior Analysis, 28*, 27–37. https://doi.org/10.1901/jaba.1995.28-27

*Lanovaz, M. J., Fletcher, S. E., & Rapp, J. T. (2009). Identifying stimuli that alter immediate and subsequent levels of vocal stereotypy: A further analysis of functionally matched stimulation. *Behavior Modification, 33*(5), 682–704. https://doi.org/10.1177/0145445509344972

*Lanovaz, M. J., & Argumedes, M. (2009). Using the three-component multiple-schedule to examine the effects of treatments on stereotypy. *Journal on Developmental Disabilities, 15*, 64–68.

*Lanovaz, M. J., & Argumedes, M. (2010). Immediate and subsequent effects of differential reinforcement of other behavior and noncontingent matched stimulation on stereotypy. *Behavioral Interventions, 25*(3), 229–238. https://doi.org/10.1002/bin.308

*Lanovaz, M. J., & Sladeczek, I. E. (2011). Vocal stereotypy in children with autism: Structural characteristics, variability, and effects of auditory stimulation. *Research in Autism Spectrum Disorders, 5*(3), 1159–1168. https://doi.org/10.1016/j.rasd.2011.01.001

Lanovaz, M. J., Rapp, J. T., & Fletcher, S. E. (2010). Expanding functional analysis of automatically reinforced behavior using a three-component multiple-schedule. *European Journal of Behavior Analysis, 11*(1), 17–27. https://doi.org/10.1080/15021149.2010.11434331

LeBlanc, L. A., Patel, M. R., & Carr, J. E. (2000). Recent advances in the assessment of aberrant behavior maintained by automatic reinforcement in individuals with developmental disabilities. *Journal of Behavior Therapy and Experimental Psychiatry, 31*(2), 137–154. https://doi.org/10.1016/s0005-7916(00)00017-3

Ledford, J. R., & Wolery, M. (2013). Procedural fidelity: An analysis of measurement and reporting practices. *Journal of Early Intervention, 35*(2), 173–193. https://doi.org/10.1177/1053815113515908

Ledford, J. R., Lane, J. D., & Gast, D. L. (2018). Dependent variables, measurement, and reliability. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 97–132). Routledge.

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

*Llinas, M., Wilder, D. A., & Walz, R. (2022). A comparison of continuous, dense, and lean schedules of noncontingent access to matched competing stimuli to reduce stereotypy. *Behavioral Interventions, 37*(3), 660–672. https://doi.org/10.1002/bin.1867

*Love, J. J., Miguel, C. F., Fernand, J. K., & LaBrie, J. K. (2012). The effects of matched stimulation and response interruption and redirection on vocal stereotypy. *Journal of Applied Behavior Analysis, 45*(3), 549–564. https://doi.org/10.1901/jaba.2012.45-549

*Luiselli, J. K. (1994). Effects of noncontingent sensory reinforcement on stereotypic behaviors in a child with post traumatic neurological impairment. *Journal of Behavior Therapy & Experimental Psychiatry, 25*, 325–330. https://doi.org/10.1016/0005-7916(94)90041-8

Moeyaert, M., Zimmerman, K. N., & Ledford, J. R. (2018). Synthesis and meta-analysis of single case research. In J. R. Ledford & D. L. Gast (Eds.),*Single case research methodology* (3rd ed., pp. 393–416). Routledge.

Mok, S. Y., Rupp, D., & Holzberger, D. (2023). What kind of individual support activities in interventions foster pre-service and beginning teachers' self-efficacy? A meta-analysis. *Educational Research Review, 40*, Article 100552. https://doi.org/10.1016/j.edurev.2023.100552

Muharib, R., Dowdy, A., Rajaraman, A., & Jessel, J. (2022). Contingency-based delay to reinforcement following functional communication training for autistic individuals: A multilevel meta-analysis. *Autism, 26*(4), 761–781. https://doi.org/10.1177/13623613211065540

Northup, J., Wacker, D., Sasso, G., Steege, M., Cigrand, K., Cook, J., & DeRaad, A. (1991). A brief functional analysis of aggressive and alternative behavior in an outclinic setting. *Journal of Applied Behavior Analysis, 24*(3), 509–522. https://doi.org/10.1901/jaba.1991.24-509

Nuernberger, J. E., Vargo, K. K., & Ringdahl, J. E. (2013). An application of differential reinforcement of other behavior and self-monitoring to address repetitive behavior. *Journal of Developmental & Physical Disabilities, 25*(1), 105–117. https://doi.org/10.1007/s10882-012-9309-x

Oliver, A. C., Pratt, L. A., & Normand, M. P. (2015). A survey of functional behavior assessment methods used by behavior analysts in practice: Functional behavior assessment survey. *Journal of Applied Behavior Analysis, 48*(4), 817–829.

*Piazza, C. C., Fisher, W. W., Hanley, G. P., LeBlanc, L. A., Worsdell, A. S., Lindauer, S. E., & Keeney, K. M. (1998). Treatment of pica through multiple analyses of its reinforcing functions. *Journal of Applied Behavior Analysis, 31*, 165–189. https://doi.org/10.1901/jaba.1998.31-165

*Piazza, C. C., Adelinis, J. D., Hanley, G. P., Goh, H.-L., & Delia, M. D. (2000). An evaluation of the effects of matched stimuli on behaviors maintained by automatic reinforcement. *Journal of Applied Behavior Analysis, 33*(1), 13–27. https://doi.org/10.1901/jaba.2000.33-13

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology, 68*, 99–112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. (2022). R Package "clubsandwich." https://cran.r-project.org/web/packages/clubSandwich/clubSandwich.pdf

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kauffmann, D. P. Hallahan, & J. P. Pullen (Eds.), *Handbook of special education* (2nd ed.). Routledge.

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science, 23*, 425–438.

Pustejovsky, J. E., Chen, M., Swan, D. M., & Grekov, P. (2023). SingleCaseES (Version 0.7.2) [R package]. https://doi.org/10.32614/CRAN.package.SingleCaseES

*Rapp, J. T. (2006). Toward an empirical method for identifying matched stimulation for automatically reinforced behavior: A preliminary investigation. *Journal of Applied Behavior Analysis, 39*(1), 137–140. https://doi.org/10.1901/jaba.2006.37-05

*Rapp, J. T. (2007). Further evaluation of methods to identify matched stimulation. *Journal of Applied Behavior Analysis, 40*(1), 73–88. https://doi.org/10.1901/jaba.2007.142-05

*Rapp, J. T., Vollmer, T. R., Dozier, C. L., St. Peter, C., & Cotnoir, N. (2004). Analysis of response reallocation in individuals with multiple forms of stereotyped behavior. *Journal of Applied Behavior Analysis, 37*, 481–501. https://doi.org/10.1901/jaba.2004.37-481

*Rapp, J. T., Swanson, G., Sheridan, S. M., Enloe, K. A., Maltese, D., Sennott, L. A., Shrader, L., Carroll, R. A., Richling, S. M., Long, E. S., & Lanovaz, M. J. (2013). Immediate and subsequent effects of matched and unmatched stimuli on targeted vocal stereotypy and untargeted motor stereotypy. *Behavior Modification, 37*(4), 543–567. https://doi.org/10.1177/0145445512461650

*Rapp, J. T., Cook, J. L., McHugh, C., & Mann, K. R. (2017). Decreasing stereotypy using NCR and DRO with functionally matched stimulation: Effects on targeted and non-targeted stereotypy. *Behavior Modification, 41*(1), 45–83. https://doi.org/10.1177/0145445516652370

Rincover, A. (1978). Sensory extinction: A procedure for eliminating self-stimulatory behavior in developmentally disabled children. *Journal of Abnormal Child Psychology, 6*(3), 299–310. https://doi.org/10.1007/BF00924733

*Rincover, A., Cook, R., Peoples, A., & Packard, D. (1979). Sensory extinction and sensory reinforcement principles for programming multiple adaptive behavior change. *Journal of Applied Behavior Analysis, 12*, 221–233. https://doi.org/10.1901/jaba.1979.12-221

Ringdahl, J. E., Vollmer, T. R., Marcus, B. A., & Roane, H. S. (1997). An analogue evaluation of environmental enrichment: The role of stimulus preference. *Journal of Applied Behavior Analysis, 30*(2), 203–216. https://doi.org/10.1901/jaba.1997.30-203

Rooker, G. W., Jessel, J., Kurtz, P. F., & Hagopian, L. P. (2013). Functional communication training with and without alternative reinforcement and punishment: An analysis of 58 applications. *Journal of Applied Behavior Analysis, 46*(4), 708–722. https://doi.org/10.1002/jaba.76

Rooker, G. W., Bonner, A. C., Dillon, C. M., & Zarcone, J. R. (2018). Behavioral treatment of automatically reinforced SIB: 1982–2015. *Journal of Applied Behavior Analysis, 51*(4), 974–997. https://doi.org/10.1002/jaba.492

Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 46*(1), 181–198. https://doi.org/10.1002/jaba.14

*Saylor, S., Sidener, T. M., Reeve, S. A., Fetherston, A., & Progar, P. R. (2012). Effects of three types of noncontingent auditory stimulation on vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis, 45*, 185–190. https://doi.org/10.1901/jaba.2012.45-185

Schoenen, E. C., Alsip, B. S., Martinez, J. C., Grekov, P., Aspiranti, K. B., & Hulac, D. M. (2024). A meta-analysis of fidget devices as academic and behavioral interventions. *School Psychology Review, 54*(3), 315–327. https://doi.org/10.1080/2372966x.2024.2411576

*Sidener, T. M., Carr, J. E., & Firth, A. M. (2005). Superimposition and withholding of edible consequences as treatment for automatically reinforced stereotypy. *Journal of Applied Behavior Analysis, 38*, 121–124. https://doi.org/10.1901/jaba.2005.58-04

*Slocum, S. K., Yatros, N., & Scheithauer, M. (2021). Developing a treatment for hand-clapping maintained by automatic reinforcement using sensory analysis, noncontingent reinforcement, and thinning. *Behavioral Interventions, 36*(1), 228–241. https://doi.org/10.1002/bin.1749

*Sprague, J., Holland, K., & Thomas, K. (1997). The effect of noncontingent sensory reinforcement, contingent sensory reinforcement, and response interruption on stereotypical and self-injurious behavior. *Research in Developmental Disabilities, 18*, 61–77. https://doi.org/10.1016/s0891-4222(96)00038-8

*Taylor, B. A., Hoch, H., & Weissman, M. (2005). The analysis and treatment of vocal stereotypy in a child with autism. *Behavioral Interventions, 20*(4), 239–253. https://doi.org/10.1002/bin.200

Tiger, J. H., Hanley, G. P., & Bruzek, J. (2008). Functional communication training: A review and practical guide. *Behavior Analysis in Practice, 1*(1), 16–23. https://doi.org/10.1007/bf03391716

Tincani, M., & Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science, 42*(1), 59–75. https://doi.org/10.1007/s40614-019-00191-5

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*(2), 196–209. https://doi.org/10.1037/h0100613

Vaughan, M. E., & Michael, J. L. (1982). Automatic reinforcement: An important but ignored concept. *Behaviorism, 10*(2), 217–227.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Vollmer, T. R. (1994). The concept of automatic reinforcement: Implications for behavioral research in developmental disabilities. *Research in Developmental Disabilities, 15*(3), 187–207. https://doi.org/10.1016/0891-4222(94)90011-6

Vollmer, T. R., Marcus, B. A., Ringdahl, J. E., & Roane, H. S. (1995). Progressing from brief assessments to extended experimental analyses in the evaluation of aberrant behavior. *Journal of Applied Behavior Analysis, 28*(4), 561–576.

Wolery, M., & Ezell, H. K. (1993). Subject descriptions and single-subject research. Journal of Learning Disabilities, 26(10), 642–647. https://doi.org/10.1177/002221949302601001

*Zawoyski, A. M., Bosch, A., Vollmer, T. R., & Walker, S. F. (2014). Evaluating the effects of matched and unmatched stimuli on nail biting in typically developing children. *Behavior Modification, 38*(3), 428–447. https://doi.org/10.1177/0145445514541615