A Systematic Review of Instructional Comparisons in Single Case Research

Jennifer R. Ledford, PhD, Jennifer.ledford@vanderbilt.edu
Kate T. Chazin, MEd, kate.chazin@vanderbilt.edu
Kari L. Gagnon, BA, kari.l.gagnon@vanderbilt.edu
Anne K. Lord, BA, anne.k.lord@vanderbilt.edu
Virginia R. Turner, BA, virginia.r.turner@vanderbilt.edu

Department of Special Education
Vanderbilt University
Peabody Box 228
Nashville, TN 32703

Kathleen N. Zimmerman, PhD, kathleen.zimmerman@ku.edu

Department of Special Education
University of Kansas
1122 West Campus Rd
Lawrence, KS 66045

The final version of this paper has been accepted to Remedial and Special Education. This manuscript (2/15/2019) is not the final version of record and may differ from the final, authoritative version.

Abstract

Comparison studies conducted to determine which instructional interventions are most efficient for teaching discrete behaviors to individuals with disabilities are potentially valuable, although some threats to internal validity may be more likely in these studies. Commonly-used standards do not address specific strategies that improve internal validity in comparison studies, such as counterbalancing assignment of stimuli or use of a control set. In this review, instructional comparison studies are described and, for studies with sufficient rigor, outcomes are evaluated. Comparisons often included young children with autism and were frequently conducted by researchers in self-contained classroom settings. Conclusions are drawn regarding relative effectiveness of common instructional procedures and suggestions are made regarding methodological standards for comparison designs.

A Systematic Review of Instructional Comparisons in Single Case Research

A primary aim of special education research is to establish evidence-based practices (EBPs) for improving outcomes for people with disabilities (Odom et al., 2005). Toward this aim, syntheses of research have resulted in the identification of a number of focused intervention practices (e.g., response prompting procedures) for improving outcomes. After one of these practices is established as effective for reliably producing change in one or more behaviors, it is important to establish whether it is *more* effective or efficient than other interventions and to determine what variations (i.e., contextual or procedural factors; e.g., group size, fidelity of implementation) impact student learning. Some researchers have suggested that comparison research leads to inappropriate inferences (Johnston, 1988); however, given existence of a number of EBPs, a second intervention may often provide a more reasonable comparison than a no-instruction baseline condition. That is, practitioners may often be more interested in the question “Will this work *better than the alternatives?*” rather than simply “Will this work?” Thus, comparison designs are critical for interpreting relative effectiveness of EBPs.

Much educational program planning includes goals related to a variety of discrete, non-reversible behaviors. These behaviors are often related to state or national standards but can also be related to functional outcomes—some common examples include reading words and answering factual questions (Ledford, Lane, Elam, & Wolery, 2012). Despite the importance of teaching these behaviors, relatively few single case studies are published focusing on measurement of non-reversible behaviors, and even fewer of these studies are comparative in nature (Hammond & Gast, 2010; Ledford, Barton, Severini, & Zimmerman, 2019).

Adapted alternating treatments designs (AATDs; Sindelar, Rosenberg, & Wilson, 1985) are single case designs appropriate for analyzing relative efficiency of instructional procedures

for improving non-reversible behaviors. In the context of these designs, it is possible to answer not only whether multiple conditions are *effective* for improving target behaviors, but also whether they are differentially *efficient*. For example, data can be collected regarding the number of sessions required for each participant to reach mastery criterion in each condition, allowing for fine-grained analysis about the extent to which one intervention results in faster acquisition than the other. The AATD is similar to the more-common alternating treatments design, but instead of rapidly alternating two or more conditions and measuring a single behavior, the AATD requires measurement of separate behaviors. Thus, different sets of equivalent behaviors are assigned to each condition—for example, if teaching word reading, five words could be assigned to Intervention A while five words of equal difficulty are assigned to Intervention B (Sindelar et al., 1985). Then, time-series data are collected for each set of behaviors before and during intervention implementation, with the two conditions being rapidly alternated (e.g., Intervention B on Monday, Intervention A on Tuesday). If criteria are reached for one instructional set and not the other, one intervention is *more effective* than the other. If mastery criteria are reached for both, but mastery is reached in fewer sessions for one, that intervention is *more efficient* than the other. If mastery is reached in a similar number of sessions, data are *undifferentiated*, and procedures are considered equally effective and efficient.

Parallel treatment designs (PTDs; Gast & Wolery, 1988) can be conceptualized as a series of time-lagged AATDs (Wolery et al., 2018). Comparisons of efficiency and effectiveness are similar for the two designs, though different methods for detecting or controlling for maturation and history threats are used in each. AATDs rely on measurement of a control set, which is a set of equally-difficult behaviors for which the participant receives no instruction. If data for these behaviors do not change, and behaviors assigned to intervention conditions do

change, it is unlikely that events outside the study are responsible (Sindelar et al., 1985).

Alternatively, PTDs rely on stable baseline data during time-lagged implementation of the intervention comparisons; that is, if behaviors that have not yet been intervened on remain stable while behaviors that are receiving either or both interventions improve, events outside the study are unlikely to be responsible for behavior change. In addition to controlling for threats, use of control sets or time-lagged implementation in comparison designs allows for measurement of “main effects” in addition to comparative outcomes (Ganz & Ayres, 2018, p. 79). Neither time-lagged implementation of the comparison condition or use of a control set is identified in widely-used contemporary standards as necessary (e.g., Council for Exceptional Children [CEC], 2014; What Works Clearinghouse [WWC], 2017).

In addition to strategies to reduce likelihood of history and maturation effects that are not identified in contemporary standards, comparison designs have additional complexities. Because two different sets of behaviors are assigned to interventions, it is critical that sets are equally-difficult to learn so that differences between conditions can be attributed to only the intervention. Authors can attempt to choose similar targets (e.g., an equal number of syllables in a word), but idiosyncratic characteristics may impact equality (Ledford, 2018)—for example, consider the words *dog* and *pig*, which could be perceived as being similarly difficult to read. However, if a child has a dog as a pet but has never seen a real pig, that difference may make the word *dog* more salient and therefore easier to learn. Another strategy for assessing equivalence is to collect baseline data to detect whether all behaviors are unknown. However, this may not be sufficient for establishing the *difficulty* of learning the unknown behaviors. In the absence of experimental analysis guiding the selection of equal behaviors (as suggested by Holcombe, Wolery, & Gast, 1994), the most effective way to control for equality is to teach the same stimuli to multiple

participants, in a counterbalanced fashion across conditions (Wolery et al., 2018). Although this does not guarantee equivalence, it does ensure that that one procedure will not be consistently identified as superior if differences are due to differential behavior difficulty.

A final related difficulty, author bias, can occur when behaviors or behavior sets are not randomly assigned to condition—this is especially problematic when counterbalancing is also not used. Randomly assigning stimuli to experimental condition ensures that behaviors authors identify as potentially easier to learn are not assigned to a favored intervention condition. This may be especially important when the authors of the comparison were also involved in the development of one of the variations being compared, although it can be difficult to ascertain when this occurs (Johnston, 1988).

It is unclear to what extent existing comparison studies adequately control for internal validity specific to these designs, since single case standards (e.g., CEC, 2014; WWC, 2017) do not explicitly address the additional concerns associated with the internal validity of comparison studies (Wolery et al., 2018). For example, WWC does not explicate the need for a control set, time-lagged comparisons, or initial baseline conditions; an AATD without any of these attributes could be considered to meet standards (assuming other criteria are met). Moreover, most currently available and widely used quantitative metrics are intended for use with reversible behaviors, and are not appropriate for comparison designs in which non-reversible behaviors are measured. For example, most metrics do not perform well when trends are present in both conditions (Moeyaert, Zimmerman, & Ledford, 2018), which is the expected data pattern when assessing two different instructional conditions for teaching discrete skills. Insufficient guidance regarding internal validity and lack of attention to appropriate metrics for comparative designs may contribute to the dearth of systematic reviews of these studies; they are further complicated

by the problems associated with syntheses such as publication bias and inability to efficiently synthesize visual analysis outcomes (Moeyaert et al., 2018; Shadish, Hedges, Horner, & Odom, 2015).

The purpose of this review is to summarize single case research comparing instructional procedures for teaching discrete behaviors to individuals with disabilities to answer both methodological questions about comparison designs and substantive questions about instructional research. Research questions were: (1) What are the characteristics of participants, settings, and implementers? (2) To what extent does the available research adequately control for threats to internal validity? (3) For comparisons that adequately control for threats to internal validity, what conclusions can be drawn about the relative efficiency of (a) different prompting procedures and (b) different instructional variations? And finally, (4) Do sources include measurement of maintenance, generalization, and consumer preference, and if so, what are the results of these assessments?

Methods

Search

A search for peer-reviewed and non-peer-reviewed sources was conducted on September, 7, 2018 using two databases: PsycINFO and Proquest Dissertations and Theses Global. The following search string was used to identify potentially eligible sources: (autis* OR retard* OR disabilit* OR delay OR handicap) AND AB(prompt* OR "time delay" OR instruction OR teaching) AND AB(compar* OR effective* OR efficien* OR relative OR superior OR fidelity OR integrity OR evaluat*) AND ("single case" OR "single subject" OR "single-case" OR "single-subject" OR "alternating treatment*" OR "parallel treatment*" OR multielement OR "repeated acquisition"). Eligibility was determined by two independent coders; agreement was

98%; differences were resolved via discrepancy discussions between coders with the first author serving as the final decision-maker when disagreements persisted. See PRISMA diagram (Moher et al., 2009) in Figure 1 for search results. A total of 5,582 potential sources were screened, 258 full text sources were reviewed for inclusion, and 98 were eligible for inclusion (3 of which were duplicate sources). To meet inclusion criteria, sources had to include at least one comparison of two or more interventions designed to teach unknown, discrete, non-reversible behaviors to individuals with disabilities. Sources had to include graphed data for visual analysis, and to confirm the behaviors were unknown via a baseline condition. To be included in the outcomes analysis, studies had to include independent and dependent reliability measurement, as well as controls for history and maturation via measurement of a control set or inclusion of time-lagged intervention comparisons. See Table 1 for criteria related to inclusion in the descriptive and outcomes analyses organized according to PICOS criteria (O'Connor, Green, & Higgins, 2008).

Coding

Descriptive coding was conducted for all eligible sources; outcomes were coded for comparisons meeting minimum internal validity requirements. Sources included in only the descriptive analysis are denoted with a single asterisk and sources included in both analyses are denoted with two asterisks (see [Supplemental References Document](#)). All descriptive and outcomes coding was conducted by two of eight independent coders (e.g., 100% double-coded); the first author, a single case design expert and doctoral level board-certified behavior analyst (BCBA-D) coded 100% of eligible sources for outcomes and intervention descriptions and at least 20% of sources in other domains. The other seven coders were assigned coding from one or two of the domains described below; these coders were graduate students in a special education program. Final codes were determined via discrepancy discussions—in the case of a

disagreement in the discussion, the first author determined the correct code. Agreement exceeded 95% for all domains with agreement per code ranging from 86.5-100%; see [Supplemental Agreement Document](#).

General codes are described below—see [Supplemental Coding Document](#) for exact categories (e.g., list of disabilities coded). Information was recorded via dichotomous 0/1 (no/yes) codes except for participant name, descriptions of comparisons (e.g., to identify which graph in a study was being coded), recording the superior intervention type by name, and codes that required a numeral (e.g., participant age). Participant coding was completed for each eligible participant in a given source ($n = 353$); settings, implementers, intervention comparison type, and secondary outcomes were coded by source ($n = 95$; with multiple codes possible for a single source), and all other codes were completed for each comparison in a source, defined as one binary comparison between two interventions ($n = 920$). Multiple comparisons for a source were coded when multiple AATDs were conducted, a PTD was conducted (with each tier constituting one comparison), or when three or more interventions were compared in a single AATD (with each intervention compared with another intervention—e.g., B vs. C, B vs. D, C vs. D).

Descriptive Coding. The following *participant* information was coded: age, gender (use of gendered pronouns was considered sufficient evidence), race, disability, and educational placement. Coders also recorded the *setting and implementer* of instructional sessions. To code teacher, teaching assistant, or therapist, an explicit indication that the individual served the participant outside of research contexts was required; otherwise use of these terms when referring to an implementer resulted in use of the “not reported” code. *Dependent variable and measurement characteristics* were also coded. The dependent variable was categorized as reading words, vocabulary or picture identification, other literacy skills, answering non-academic

factual questions, matching, math skills, other academic skills, or other. Coders also assessed the *unit of measurement* reported on the y-axis and whether *maintenance* and *generalization* were measured and reported. Coders assessed whether *social validity* was measured via implementer, stakeholder, or participant preference.

Coders assessed *internal validity* by coding whether authors collected interobserver agreement (IOA) and procedural fidelity (PF) data, whether data were collected with sufficient frequency (25% of sessions), and whether results of reliability assessments were acceptable (90%). This agreement threshold is higher than an often-cited minimum of 80% (e.g., WWC, 2017); however, as described by Ledford, Lane, and Tate (2018), agreement for trial-based discrete behaviors (the type measured in these studies) should generally be considerably higher than 80%. Coders also recorded whether baseline included at least three measurement occasions, whether baseline data were equivalent across conditions, whether authors used random assignment, whether authors counterbalanced assignment across participants, whether control data were collected and graphed, and whether there were three time-lagged introductions of the comparison condition (i.e., parallel treatments designs). Finally, coders assessed the nature of the *intervention comparison* for each source. For example, if a constant time delay procedure was compared to a system of least prompts procedure, the intervention comparison type was coded as “single prompt versus hierarchical procedure” (see [Supplemental Intervention Glossary](#) for definitions of intervention comparison terms and categorization of procedures).

Outcomes coding. The first two authors coded outcomes for all sources meeting minimum internal validity requirements. These requirements were: adequate control for history and maturation threats, and adequate evidence of independent and dependent variable reliability (see Table 1). Coders first assessed whether each comparison included adequate control for

history and maturation using visual analysis. Specifically, coders checked for clear differentiation between trends in instruction and non-instruction (baseline, control) conditions, with minimal changes in non-instruction condition over time (Ledford et al., 2019). For comparisons with adequate control, coders assessed which outcome occurred regarding effectiveness: (a) both interventions effective (improvements that occurred when and only when intervention was implemented, assessed via visual analysis, with data reaching maximum values such as 100% correct or the stated mastery criteria, when one was provided), (b) neither intervention effective, or (c) only one intervention effective. Then coders recorded whether data were differentiated (e.g., both effective but one more efficient) or undifferentiated. Finally, coders recorded via text which intervention was more effective or efficient (or indicated “neither” or “undifferentiated”). We considered differences in mastery of one session or fewer than 10% of sessions as equivocal (undifferentiated). Finally, data regarding secondary outcomes were coded, including: (a) whether generalization and maintenance results were consistently in favor of one intervention, and (b) whether consumer preference favored one intervention.

Search results. A total of 95 unique sources were identified (68 peer-reviewed articles and 27 dissertations or theses); they were published between 1986 and 2018, with 56 published since 2010. Three published articles were associated with identified unpublished dissertations or theses; these were analyzed as peer-reviewed sources (total number of sources = 98). In sum, there were 920 comparisons (see Tables 2 & 3).

Results

Participants

A total of 353 individuals, from ages 1 to 38 years, participated in the included comparisons (mean = 7 years, median = 6 years). Most participants (70%) were male, while

20% were female and 10% were unreported. Race or ethnicity was reported for only 16% of participants; when reported, most participants (61%; $n = 35$) were identified as White, 15% ($n = 8$) were identified as Black, 15% ($n = 8$) were identified as Asian, and 7% were identified as having Hispanic ethnicity ($n = 4$). No participants were identified as having Middle Eastern ethnicity and only one was identified as being multiracial. Many participants were identified as having autism spectrum disorder (ASD; 61%, $n = 217$) or intellectual disability (ID; 29%, $n = 103$); some were identified as having a speech or language impairment (10%, $n = 35$, including indication that a child received speech-language therapy), a learning disability (8%, $n = 27$), or a developmental delay (4%, $n = 15$). Only 14% of participants had neither ASD nor ID. Participants were more rarely identified with health impairment diagnoses such as seizures ($n = 13$), physical disabilities ($n = 12$), sensory disabilities (e.g., hearing loss, vision loss; $n = 9$), “multiple disabilities” ($n = 6$), attention-deficit/hyperactivity disorder (ADHD; $n = 5$), or behavioral disorders ($n = 2$). The most commonly-reported educational placement for participants was self-contained classrooms (30%, $n = 105$), with fewer participants served in general education classrooms (14%, $n = 50$) or both general education and self-contained classrooms (10%, $n = 34$). Educational placement was not reported for 166 participants.

Settings and Implementers

Setting was coded separately from educational placement. Some sources reported multiple settings (e.g., instruction was implemented in a general education classroom for one participant and a segregated classroom for another) and some participants did not participate in the study in their classroom (e.g., placement was general education, study setting was home). Authors reported settings were segregated classrooms ($n = 32$), clinics ($n = 18$), homes ($n = 16$), research or hospital settings ($n = 14$), general education classrooms ($n = 12$), or unspecified or

unreported locations ($n = 28$). The majority of sources reported researcher implementers ($n = 62$); other implementers included teachers ($n = 19$), teaching assistants ($n = 5$), or peers ($n = 1$). Four sources reported multiple implementers and implementer was unreported in 12 sources.

Dependent Variables and Measurement

The most commonly taught behaviors were identifying pictures or vocabulary ($n = 450$ comparisons, 49%), answering factual non-academic questions (including features and functions; $n = 181$; 20%), and reading words ($n = 143$, 16%). Less commonly-targeted behaviors were non-reading literacy skills ($n = 27$, 3%), math skills ($n = 18$, 2%), and other academic behaviors ($n = 59$, 6%). Behaviors categorized as “other” ($n = 35$, 4%) included play behaviors and imitation. Most comparisons included outcomes reported as a percentage correct ($n = 787$, 86%); some reported number correct ($n = 85$, 9%) or cumulative number ($n = 48$, 5%). Maintenance of behaviors was measured in 604 comparisons (66%), generalization was measured in 232 comparisons (25%), and preference for instructional procedure was measured less frequently ($n = 129$ for participants, $n = 42$ for implementers, and $n = 50$ for other stakeholders).

Intervention Type

There were 106 intervention comparisons across sources, including sources with multiple comparison types. For example, if two single prompt procedures (e.g., constant and progressive time delay) and one hierarchical procedure (e.g., system of least prompts) were included in a design, three comparisons were present (constant vs. progressive time delay, constant time delay vs. system of least prompts, and progressive time delay vs. system of least prompts), and two intervention comparison types were present (single prompt versus single prompt [$n = 1$] and single prompt versus hierarchical prompt [$n = 2$]). Common comparison types included prompting procedure comparisons ($n = 27$) and consequence ($n = 19$), antecedent ($n = 13$), and

trial arrangement ($n = 9$) variations. Others were prompt mode ($n = 8$), trial ordering ($n = 7$), group size ($n = 6$), fidelity ($n = 5$), and frequency ($n = 2$). We further divided three categories into subcategories (prompting procedures—single prompt procedure comparisons, hierarchical comparisons, single prompt versus hierarchical, and errorless versus error correction; consequence variations—corrective feedback, instructive feedback, reward procedures, and reward type; antecedent variations—different materials, presence of additional verbal or visual information, wait time differences, and other).

Internal Validity

All included comparisons reported measurement of dependent variable reliability data, with 81% of them including measurement in at least 25% of sessions and 98% reporting overall agreement of at least 90%. Fidelity measurement was also common ($n = 736$, 80%), with 67% of sources including data collected for at least 25% of sessions and nearly all ($n = 733$) reporting average fidelity of greater than 90%. Most comparisons ($n = 816$, 89%) also included three or more baseline measurement occasions. Internal validity indicators specific to comparison designs for non-reversible behaviors were less commonly reported. Slightly more than half of comparisons ($n = 464$) included behaviors that were randomly assigned to conditions while fewer counterbalanced behavior assignment across participants ($n = 256$, 28%). Nearly half of comparisons either included graphed data in a control condition or conducted time-lagged intervention comparisons with at least three tiers ($n = 447$, 49%).

Outcomes

Of 920 experimental comparisons, 408 were included in the outcomes analysis due to controls for history, maturation, instrumentation, and infidelity threats. Some included control or baseline data with increasing trends ($n = 45$) or were not interpretable due to the two conditions

having multiple differences between them (e.g., different prompting procedure and different trial arrangement; $n = 12$). Thus, we evaluated outcomes for 351 experimental comparisons. Of these comparisons, data in 161 ($n = 46\%$) showed both interventions were *effective* and one was more *efficient*, while data in 135 (38%) showed both interventions were *effective* and similarly *efficient*. Few comparisons included data that showed only one ($n = 45$, 13%) or neither ($n = 10$, 3%) intervention was *effective*. Sources that counterbalanced behaviors across participants resulted in 56% of comparisons in which one intervention was identified as more effective or efficient and 44% similarly efficient.

Tables 4 and 5 show comparison types that resulted in differentiated and undifferentiated outcomes. Outcomes are discussed below in two separate sections—prompting procedure comparisons and instructional variation comparisons. Outcomes are further described by intervention category; specific information about comparisons and which sources contributed to these conclusions can be found in the [Supplemental Outcomes Description](#) document. Below, we identify superior interventions, defined as an intervention which was shown to be more effective or efficient in the majority of comparisons; we also discuss the extent to which outcomes were *consistent* or *inconsistent* for individual participants, when the same comparison was conducted multiple times for a single participant. Consistency occurred when one intervention type resulted in faster learning for all comparisons or when all comparisons were undifferentiated.

Prompting procedure comparisons. Prompting procedure comparisons included those in which two unique prompting procedures were assessed for relative efficiency. As shown in Table 2, these comprised 131 of 351 total comparisons; 37%.

Single prompt comparisons. Three types of single prompt comparisons were evaluated: (a) constant versus progressive time delay, (b) constant time delay versus simultaneous

prompting, and (c) simultaneous prompting versus antecedent prompt and test (APT). Results by type and corresponding citations are available in the [Supplemental Outcomes Description](#) document. As shown in Table 5, comparisons (a) and (b) resulted in no identification of a superior intervention via a majority. Comparison (c) resulted in APT being identified as the superior intervention, with 11/12 comparisons resulting in identification of APT as the most efficient intervention. However, this comparison was only evaluated in a single source. Across all single prompt comparisons, when multiple comparisons were conducted with individual participants ($n = 17$ participants), 11 were inconsistent (e.g., one intervention was superior in the first comparison and the other was superior in the second comparison); 2 comparisons included consistent undifferentiated data; the remaining 4 were consistently in favor of one intervention.

Hierarchical comparisons. There were fewer hierarchical comparisons; in these comparisons, which included system of least prompts, most-to-least prompts, and flexible prompt fading, no superior interventions were identified. For all four participants with multiple comparisons, inconsistent outcomes were identified, with all including at least one undifferentiated comparison.

Single prompt versus hierarchical comparisons. When single prompt and hierarchical procedures were compared, data were undifferentiated in a majority of comparisons ($n = 17$); in 14 comparisons that were differentiated, single prompt procedures were more efficient in 11. When multiple comparisons within a participant existed, several were consistently undifferentiated ($n = 2$), but most were inconsistent ($n = 7$), with four of the inconsistent findings resulting in initial efficiency in favor of the single prompt procedure and later comparisons resulting in undifferentiated data.

Errorless versus error correction comparisons. When errorless and error correction procedures were compared, many resulted in undifferentiated data ($n = 13$) or identification of error correction as superior ($n = 13$), with fewer identifying errorless procedures as more efficient ($n = 4$). Nine participants had multiple comparisons, with most resulting in inconsistent outcomes ($n = 5$) and 7 of 9 including at least one undifferentiated comparison.

Instructional variation comparisons. Instructional variation comparisons included those in which the same instructional procedures were used (e.g., a single prompting procedure), although one component of the instruction varied across conditions. As shown in Table 3, these comprised 220 of 351 total comparisons; 63%. Results by type of comparison, along with corresponding citations, are available in the [Supplemental Outcomes Description](#) document. Below, outcomes are described for specific comparison types that included at least 10 comparisons across at least two sources—these include (a) fidelity of instruction, (b) group size, (c) massed versus embedded trials, and (d) presence or absence of instructive feedback.

Fidelity comparisons. When higher- and lower-fidelity instruction was compared, more efficient learning most often occurred during higher-fidelity conditions ($n = 25$) or data were undifferentiated ($n = 19$), with lower-fidelity instruction less commonly resulting in more efficient learning ($n = 7$). Nine comparisons in this group were between two different low fidelity variations. Two participants with multiple fidelity comparisons consistently learned faster in higher-fidelity conditions; the remaining three participants had inconsistent outcomes.

Group size comparisons. Data were undifferentiated in nearly half of comparisons ($n = 16$), while individual instruction resulted in more efficient learning in most of the remaining comparisons ($n = 14$), and group instruction resulted in more efficient learning in others ($n = 6$). Nine participants had multiple comparisons; one each had consistent results in favor of

individual and group instruction; remaining comparisons were inconsistent, including four in which participants initially learned better in one condition but later data were undifferentiated.

Trial arrangement comparisons. Results were inconsistent, with approximately equal numbers of comparisons finding massed ($n = 5$) or embedded ($n = 3$) instruction more efficient, and 3 additional comparisons resulting in undifferentiated data. Only one comparison was conducted multiple times for a single participant; it resulted in inconsistent outcomes.

Instructive feedback comparisons. Results were inconsistent, with 5 comparisons resulting in undifferentiated data and most of remaining comparisons resulting in more efficient learning when instructive feedback was present ($n = 5$) versus absent ($n = 2$). Five participants had multiple instructive feedback comparisons; all included inconsistent results—in four, participants learned faster with instructive feedback in the initial comparison and had undifferentiated data in subsequent comparisons.

Secondary variables. We assessed differential maintenance, generalization, and consumer preference as secondary outcomes. Maintenance was measured and reported in 31 sources—we considered maintenance to have occurred if data were within 75% of the stated mastery criterion (or optimal levels if no criteria were named). In 18 sources, behaviors assigned to both conditions maintained; in 12, there were mixed results (e.g., maintenance occurred in only one condition for one comparison and for the other condition in a second comparison); and in 1 source, behaviors assigned to neither intervention were reported to have maintained. No sources reported consistent differential maintenance. Generalization was measured and reported in 10 sources—we considered generalization to have occurred if data were within 75% of the stated mastery criterion (or optimal levels if no criteria were named). In 2 sources, behaviors assigned to both conditions generalized; and in 8, there were mixed results (e.g., generalization

occurred in only one condition for one comparison and for the other condition in a second comparison). No sources reported consistent differential generalization. Consumer preference was measured and reported in 12 sources. In nine sources, results were mixed across respondents; in one source, respondents reported having no preference. In two sources, a consistent preference was obtained—simultaneous prompting was preferred over constant time delay and choice of rewards was preferred over no choice.

Discussion

The purpose of this paper was to review comparative research on intervention strategies for teaching discrete behaviors to individuals with disabilities. We included unpublished dissertations in an attempt to reduce the likelihood of publication bias (Shadish, Zelinsky, Vevea, & Kratochwill, 2016), especially given the a priori expectation that undifferentiated results were likely to be prevalent. Although the number of included studies ($n = 95$), participants ($n = 353$), and comparisons ($n = 920$) was large, the comparisons were heterogeneous and relatively few of the comparisons met minimal internal validity requirements for assessing outcomes. Results of the review should be considered in light of several notable limitations. First, no effect sizes were calculated, given the inappropriateness of available metrics for appropriately capturing data for non-reversible behaviors in comparison designs. Additional work is needed in this area. Second, some types of comparisons were assessed in a single source; replications are needed across categories. Third, only discrete behaviors were included in the review, per the inclusion criteria—conclusions only apply to this type of behavior. Fourth, additional search strategies such as forward searches may have resulted in additional sources; we did not conduct these searches because we hypothesized sources citing included publications likely matched the publication on content rather than design type (i.e., may have been unlikely to include

comparison designs). Thus, additional relevant studies may exist that were not included in this review. Finally, differences of 10% were considered equivocal. This decision was somewhat arbitrary—we discuss implications of this decision and suggestions for future research in the methodological conclusions section. Despite limitations, a number of important conclusions can be drawn regarding both substantive content and methodology.

Substantive Conclusions, Implications, and Future Directions

Participants, settings, and implementers. Participants in the included sources tended to be young and to be identified as having ASD or ID, similar to other reviews specific to instructional strategies (Morse & Schuster, 2004; Shepley, Lane, & Ault, 2018). This may be because young children are often taught discrete behaviors, while older children may be taught more complex skills that build on these early learning targets; students with ASD and ID may have been included because they, in particular, benefit from systematic instruction for learning discrete behaviors (Thompson, 2011). Participants were most often served in self-contained settings and studies most often took place in separate settings (self-contained classrooms, clinics, hospitals); this is similar to the findings from other reviews (e.g., Ledford et al., 2012; Morse & Schuster, 2004). This may suggest that conclusions from this review are primarily applicable to young children with more intensive support needs, in segregated settings. Additional research is needed for older children with a variety of disabilities in inclusive settings. Few authors reported race; previous reviews have not included information about participant race. This lack of reporting limits the extent to which we can identify for whom these comparisons have been conducted. Most studies were conducted with researcher implementers, a finding consistent with other reviews of special education instruction research (Ledford et al., 2012; Shepley et al., 2018); additional data are needed to determine to what extent outcomes are similar with

indigenous implementers and to what extent these implementers have differential preferences for various procedures.

Independent and dependent variables. Similar to previous reviews (e.g., Ledford et al., 2012), students were most often taught to name pictures or other stimuli, read words, or answer factual questions. Common comparisons included different prompting procedures, high versus low fidelity implementation, presence versus absence of instructive feedback, group size, and trial arrangements. Most intervention comparison types were assessed in only a few adequately rigorous sources.

Outcomes. Nearly all comparisons resulted in mastery-level learning by participants for both instructional conditions; this finding is less likely to be influenced by publication restrictions given that the search included both published and unpublished sources. That is, the review supports the notion that systematic instruction, regardless of procedure or variation type, results in mastery. Although more than half of outcomes were differentiated, consistent findings within participants and studies was uncommon. We draw a number of recommendations based on comparisons that include multiple comparisons and sources (see Table 6). Because counterbalancing was rare, results must be interpreted cautiously and as potentially related to differential behavior difficulty—additional replications are needed. It is also important that inconsistent results within participant may point to their ability to “learn to learn”—that is, participants with disabilities may initially learn faster given one procedure but quickly learn how to efficiently learn from a variety of procedures given that they are all systematic and implemented as planned. This phenomenon may result in increasingly undifferentiated learning over time. Additionally, given that instructional procedures may result in undifferentiated outcomes often, authors of comparison studies should always measure participant preference;

given equal efficiency, this measure of social validity is of paramount concern. Although it interferes with the ability of authors to randomize session order, ongoing preference data can be collected in the context of rapidly alternating conditions by including the provision of child choice (e.g., Ledford, Chazin, Harbin & Ward, 2017).

Regarding instructional variations, practitioners should (a) provide corrective feedback when using prompting procedures, given that it is likely to increase learning efficiency; (b) provide instructive feedback, given that it increases the number of targets a child can learn while either increasing or maintaining efficiency for primary targets; and (c) conduct prompting procedures with high fidelity. They might also consider using small group rather than individual instruction, given that small group instruction was not generally associated with slower acquisition and it is associated with other benefits (e.g., less segregation, observational learning opportunities). More work is needed to draw confident conclusions about group size, including whether children “learn to learn” efficiently across group sizes and for group sizes larger than dyads and triads. Because no conclusive evidence was collected in favor of either massed trials or trials embedded during other activities, and given that one study showed preference might be related to rate of acquisition (Ledford et al., 2017), practitioners should use child and implementer preference to govern choice; then monitor progress and use a different trial arrangement if learning does not occur or acquisition is slow.

Practitioners should use systematic prompting procedures, as nearly all were associated with acquisition of a range of behaviors (consistent with previous reviews for single prompt procedures but not for reviews of hierarchical procedures; Shepley et al., 2018; Walker, 2008). When single prompt and hierarchical procedures resulted in differentiated data, single prompt procedures were more often effective. Interestingly, when errorless procedures were compared

with error correction procedures, many comparisons showed participants learned only with the error correction procedure—given the large body of literature showing errorless procedures to be effective, additional research is needed to determine in which contexts they may be ineffective. This finding may be specific to situations contrived by comparison designs—that is, when errorless and non-errorless procedures are alternated, the child is expected to wait for a prompt during some sessions and is expected to guess and receive corrective feedback for other sessions. This might result in confusion and increased rates of errors in errorless conditions compared with studies when these conditions are implemented in the absence of the error-correction comparison. Given differences between prompting procedures were inconsistent, consider selection of procedure based on child and implementer preference and likelihood of use.

Because fidelity was consistently high in this review and because researchers were the most common implementer, conclusions about feasibility under typical constraints (e.g., teachers in classrooms) are not possible. Additional research on ease of use and likely fidelity with indigenous implementers is warranted.

Methodological Conclusions, Implications, and Future Directions

Although inclusion in the descriptive analysis did not require it, nearly all studies measured dependent variable reliability that met or exceeded current standards (25% of sessions with 90% or greater agreement). Most also measured procedural fidelity for 25% of sessions with 90% or greater fidelity. Despite high rigor according to typical single case standards, authors less often reported controlling for threats specific to comparison designs and measurement of non-reversible behaviors. We chose not to evaluate outcomes for studies that did not include controls for history and maturation (i.e., a control set or time-lagged implementation)—this limits the conclusions that can be drawn from the larger body of research

but increases confidence that results obtained were not due to internal validity threats rather than true differences between conditions. However, differences in behavior difficulty and potential author bias may have resulted in unfair comparisons; few studies included controls for these potential problems despite the fact that they have been long-associated with comparison designs (Johnston, 1988). Based on our findings, we recommend two contemporary standards with regards to comparison designs for assessing differential acquisition of non-reversible behaviors.

Standard #1: Include a control set for all comparisons, or use time-lagged implementation with at least three tiers.

Standard #2: Counterbalance assignment of behavior sets to conditions across participants. When counterbalancing is not possible, authors should randomly assign behavior sets to intervention conditions; this limits potential for bias but may not result in confidence that differential outcomes were due to intervention differences rather than unequal behaviors sets.

In addition to these standards, we make two recommendations regarding future instructional comparisons research. First, we hypothesize that a change in common practice regarding measurement may lead to more meaningful conclusions when AATDs are used. We made an arbitrary decision to code differences of at least 10% as “differentiated”. In practice, a 10% increase in learning over a long period of time (e.g., a school year) could yield meaningful differences. However, additional research is needed to determine to what extent use of different procedures impacts efficiency over a sustained period of time. Small differences in efficiency might be established over time by using AATDs while measuring the *cumulative number* of behaviors mastered—for example, given a pool of stimuli, practitioners or researchers might assign a large number to each of two sets and measure the cumulative number mastered over a long period of time. This type of assessment may take longer than the typical AATD (e.g.,

assigning only a few stimuli to each of two conditions and measuring percentage correct), but may allow for analyses that more closely align with practical considerations. This type of measurement was rare in included studies, and no studies that used cumulative number as a dependent variable met internal validity requirements for inclusion in the outcomes analysis.

In addition, extensions of instructional research may include the purposeful selection of participants who might be theoretically assumed to benefit from one variation over another and to conduct comparison across groups. For example, researchers have surmised that participants who are readily imitative of peers might benefit from group rather than individual instruction and participants with repetitive play might benefit more from massed trial instruction rather than instruction embedded in play (Ledford, Lane, & Barton, 2019). In our review, we did not find any studies in which authors explicated research questions about the impact of participant characteristics on differential acquisition of learning. Thus, we make two recommendations for future research.

Recommendation #1: Conduct comparisons with large sets of behaviors while measuring cumulative mastery over time.

Recommendation #2: Conduct studies comparing response to intervention across groups, with explicated a priori research questions.

Finally, it bears repeating—practitioners should monitor progress to identify whether the EBPs they use are resulting in expected gains. Given inconsistency in results across comparison types, it is likely that some variations will result in faster learning for some individuals and not others—thus, practitioners should be prepared by familiarizing themselves with EBPs, potential instructional variations, and the tools needed to analyze differential behavior acquisition.

References

- Council for Exceptional Children. (2014). Council for Exceptional Children standards for evidence-based practices in special education. Retrieved from:
<https://www.cec.sped.org/Standards/Evidence-Based-Practice-Resources-Original>
- Ganz, J. B., & Ayres, K. M. (2018). Methodological standards in single-case experimental design: Raising the bar. *Research in Developmental Disabilities, 79*, 3-9. doi: 10.1016/j.ridd.2018.03.003
- Gast, D. L., & Wolery, M. (1988). Parallel treatments design: A nested single subject design for comparing instructional procedures. *Education and Treatment of Children, 270-285*.
- Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single case research designs: 1983–2007. *Education and Training in Autism and Developmental Disabilities, 45*, 187-202.
- Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single-subject research: Description of designs and discussion of problems. *Topics in Early Childhood Special Education, 14*(1), 119-145. doi: 10.1177/027112149401400111
- Johnston, J. M. (1988). Strategic and tactical limits of comparison studies. *The Behavior Analyst, 11*, 1-9. doi: 10.1007/BF03392448
- Ledford, J. R. (2018). No randomization? No problem: Experimental control and random assignment in single case research. *American Journal of Evaluation, 39*, 71-90. Doi: 10.1177/1098214017723110
- Ledford, J. R., Barton, E. E., Severini, K. E., & Zimmerman, K. N. (2019). A primer on single-case research designs: Contemporary use and analysis. *American Journal on Intellectual and Developmental Disabilities, 124*, 35-56
- Ledford, J. R., Chazin, K. T., Harbin, E. R., & Ward, S. E. (2017). Massed trials versus trials

- embedded into game play: Child outcomes and preference. *Topics in Early Childhood Special Education*, 37, 107-120. doi: 10.1177/0271121417704703
- Ledford, J. R., Lane, J. D., & Barton, E. E. (2019). *Methods for teaching in early education*. New York, NY: Routledge.
- Ledford, J. R., Lane, J. D., Elam, K., & Wolery, M. (2012). Using response prompting procedures during small group direct instruction: Outcomes and procedural variations. *American Journal of Intellectual and Developmental Disabilities*, 117, 413-434. doi: 10.1352/1944-7558-117.5.413
- Ledford, J. R., Lane, R., & Tate, R. (2018). Evaluating quality and rigor in single case research (p. 365-392). In J. R. Ledford & D. L. Gast (Eds.). *Single case research methodology. Applications in special education and behavioral sciences*. New York: Routledge.
- 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*, New York, New York: Routledge.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med*, 6, e1000097. Doi: 10.1371/journal.pmed.1000097
- Morse, T. E., & Schuster, J. W. (2004). Simultaneous prompting: A review of the literature. *Education and Training in Developmental Disabilities*, 39, 153-168.
- O'Connor, D., Green, S., & Higgins, J. (2008). Defining the review question and developing criteria for including studies. In: J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0*. The Cochrane Collaboration.

- Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, 71, 137-148. doi: 10.1177/001440290507100201
- Shadish, W.R., Hedges, L.V., Horner, R.H., and Odom, S.L. (2015). The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research (NCER 2015-002) Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- Shadish, W. R., Zelinsky, N. A., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis*, 49, 656-673.
- Shepley, C., Lane, J. D., & Ault, M. J. (2018). A review and critical examination of the system of least prompts. *Remedial and Special Education*. doi: 10.1177/0741932517751213
- Sindelar, P. T., Rosenberg, M. S., & Wilson, R. J. (1985). An adapted alternating treatments design for instructional research. *Education and Treatment of Children*, 8, 67-76.
- Thompson, T. (2011). Individualized autism intervention for young children. Discrete trial and naturalistic strategies. Baltimore, MD: Brookes.
- Walker, G. (2008). Constant and progressive time delay procedures for teaching children with autism: A literature review. *Journal of Autism and Developmental Disorders*, 38, 261-275. doi: 10.1007/s10803-007-0390-4
- What Works Clearinghouse, Institute of Education Sciences. (2017). *Procedures handbook, version 4.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_procedures_handbook_v4_draft.pdf
- Wolery, M., Gast, D. L., & Ledford, J. R. (2018). Comparative designs. In J. R. Ledford & D. L. Gast (Eds.). *Single case research methodology. Applications in special education and behavioral sciences*. New York, NY: Routledge.

Table 1. Criteria for Inclusion in Descriptive Review and Outcomes Analysis

Category	Inclusion	Exclusion
Inclusion in Descriptive Review		
Participants	Individuals identified as having a disability or special education eligibility, of any age.	Individuals identified as not having a disability, including those at-risk for academic failure.
Intervention	Any intervention designed to improve acquisition of specific, targeted discrete behaviors.	Comparisons of instructional mode or implementers with identical instructional procedures (e.g., tablet versus notecard; teacher versus peer); comparison of interventions designed to teach different <i>behaviors</i> (e.g., speech versus sign). Interventions with different dosages or frequencies, if these were not intended differences to be compared.
Comparison	Second intervention condition; time-series data; initial baseline data collected for each intervention condition. Designs in which different stimuli are assigned to each condition and measured repeatedly. Comparisons must be concurrent, with no more than two data points in a row until mastery.	Comparisons between intervention and non-intervention (baseline) conditions, pre and post comparisons.

Outcomes	Non-reversible discrete behaviors presented in graphic format to allow for visual analysis.	Chained behaviors (writing, spelling), generalized academic behaviors, reversible behaviors; behaviors measured in the context of a single case design without graphic displays; rate; requesting.
Setting	No setting restrictions.	N/A

Inclusion in Outcomes Analysis

Internal Validity	Appropriate control for potential threats, defined as both (a) use of a control set or time-lagged implementation with three separate start points (e.g., parallel treatments design), (b) no indication of increasing trends in control data, and (c) collection of IOA and PF data, with acceptable results, defined as at least 90% agreement (IOA) or 90% correct implementation (PF).	Collection but not reporting of IOA or PF data; designs without control sets that also did not include at least three time-lagged intervention start points. Note that time lag of two tiers could result in presence of control data for first tier if second tier intervention began after mastery was met for first tier (e.g., baseline data for second tier served as control set for first tier).
-------------------	--	---

Note: All inclusion criteria were applied at the design rather than the source level; that is, given two experimental designs in a single publication, one could be included and one could be excluded.

IOA=interobserver agreement or dependent variable reliability. PF=procedural fidelity.

Table 2. Sources Included in Descriptive Review and Outcomes Analysis

Category	Subcategory	Descriptive Analysis	Outcomes Analysis
Prompting	Single Prompt	11 (93)	6 (56)
Procedure	Hierarchical	3 (27)	2 (14)
	Single Prompt vs. Hierarchical	4 (37)	4 (31)
	Errorless vs. Error Correction	9 (75)	4 (30)

Note: Initial numerals in each cell refer to the total number of sources. Numerals in parentheses refer to the total number of comparisons. Comparisons met minimum internal validity requirements to be included in outcomes analysis.

Table 3. Instructional Variation Comparisons included in Descriptive Review and Outcomes Analysis

Category	Subcategory	Descriptive Analysis	Outcomes Analysis
Antecedent	Materials Differences	2 (8)	0
Variation	Additional Verbal Information	5 (30)	1 (5)
	Additional Visual Information	3 (16)	1 (4)
	Varied Wait Time	2 (7)	0
	Other	1 (15)	0
Consequence	Error Correction Type	6 (102)	2 (33)
Variation	Instructive Feedback	5 (33)	2 (12)
	Reward Procedures	4 (38)	3 (19)
	Reward Type	4 (72)	2 (15)
Fidelity		5 (72)	2 (60)
Frequency		2 (6)	2 (5)
Group Size		6 (45)	3 (36)
Prompt Mode		8 (40)	1 (5)
Trial Arrangement		9 (48)	4 (20)
Trial Ordering		7 (112)	1 (6)

Note: Initial numerals in each cell refer to the total number of sources. Numerals in parentheses refer to the total number of comparisons. Comparisons met minimum internal validity requirements to be included in outcomes analysis.

Table 4. Outcome Results for Comparisons in which the Majority of Results were Differentiated.

Category	Description	Superior Intervention
Extra visual	Superimposed image versus none	No extra visual information
Error Correction (EC) Type	1. Corrective feedback v. none 2. Multiple corrections v. single corrections 3. Model plus trial as EC v. model alone	1. Corrective feedback 2. Multiple corrections 3. Model plus new trial
Instructive Feedback (IF)	Instructive feedback provided during trial v. none	Undifferentiated, instructive feedback (tie)
Reward Procedures	1. Independent v. interdependent contingency 2. Shorter v. longer delays to reinforcement 3. Choice of reward v. no choice	1. None. 2. Shorter delay 3. Choice
Reward Type	1. Praise v. praise plus other rewards. 2. Shorter v. longer delays to reinforcement 3. Choice of reward v. no choice	1. Undifferentiated. 2. Shorter delay 3. Choice
Fidelity	1. High fidelity v. low fidelity instruction 2. Errors in instruction v. prompt v. reward	1. High fidelity 2. None
Trial Ordering	Unknown trials only v. known trials interspersed	No interspersal

Note: Superior intervention = more than half of comparisons favored this intervention. Undifferentiated = more than half of comparisons were not differentiated. None = there was no majority (e.g., some comparisons resulted in faster learning in one condition, some in the other condition, and some undifferentiated).

Table 5. Outcome Results for Comparisons in which the Majority of Results were Undifferentiated

Category	Description	Superior Intervention
Single Prompt Comparisons	1. Constant versus progressive time delay	1. None
	2. Constant time delay versus simultaneous prompting	2. Undifferentiated
	3. Simultaneous prompting versus antecedent prompt and test (APT)	3. APT
Hierarchical Comparisons	Procedures that reduce prompt intrusiveness over time (most-to-least) versus those that alter intrusiveness within a session	None
Single Prompt vs. Hierarchical	Single prompt procedures (use controlling prompt only) versus hierarchical procedures (use controlling and non-controlling prompts)	Undifferentiated
Errorless vs. Error Correction	Errorless prompting versus procedures that rely on error correction to teach correct responding (i.e., error correction, no-no prompting)	None
Extra verbal	Verbal information provided v. none	Undifferentiated
Frequency	More frequent v. less frequent instruction	Undifferentiated
Group Size	Individual v. small group instruction	Undifferentiated
Prompt Mode	Direct v. indirect verbal prompts	Undifferentiated
Trial Arrangement	1. Massed v. embedded trials	1. None
	2. Individual v. group attending cues	2. Undifferentiated
	3. Shorter v. longer interval between trials	3. Undifferentiated

Note: Superior intervention = more than half of comparisons favored this intervention. Undifferentiated = more than half of comparisons not differentiated. None = there was no majority (e.g., some comparisons resulted in faster learning in one condition, some in the other condition, and some undifferentiated).

Table 6. Conclusions drawn from intervention comparisons.

Category	Conclusions
Extra Verbal Information	Extra verbal information in discriminative stimuli or prompts may vary across individuals or have no differential impact.
Extra visual	Extra visual information in the form of superimposed images may result in slower acquisition of reading words.
Corrective Feedback Type	Provision of error correction may improve efficiency of prompting procedures. Using multiple corrections (requiring a prompted response) may result in more efficient learning than using single corrections. Conducting a new trial contingent on an error, rather than requiring a corrected response, may result in more efficient learning.
Instructive Feedback (IF)	Provision of instructive feedback does <i>not</i> result in slower acquisition of targeted behaviors.
Reward Procedures	Shorter delays to reinforcement following responding increased learning efficiency. Provision of choice of reward resulted in increased learning efficiency.
Reward Type	Effectiveness of other rewards in addition to praise may vary across individuals or have no differential impact. Shorter delays to reinforcement following responding may increase learning efficiency. Provision of choice of reward may result in increased learning efficiency.
Fidelity	High fidelity instruction may result in faster learning than low fidelity instruction.
Frequency	More versus less frequent instruction may vary across individuals or have no differential impact.

Group Size	Individual versus small group instruction may vary across individuals or have no differential impact.
Prompt Mode	Impact of specific prompt mode may vary across individuals or have no differential impact.
Single Prompt	CTD and SP may vary across individuals or have no differential impact.
Comparisons	Conducting probe trials following instruction (APT) rather than prior to instruction (SP) each day may result in more efficient learning.
Hierarchical Comparisons	No clear conclusions.
Single Prompt vs. Hierarchical	Impact of single prompt versus hierarchical procedures may vary across individuals or have no differential impact. When one was more superior, it was more often single prompt; additional research is needed to determine whether this difference is consistent or whether differences can be attributed to differential behavior difficulty.
Errorless vs. Error Correction	No clear conclusions; inconsistent outcomes may be due to between-student variability but may also be attributable to differential behavior difficulty.
Trial Arrangement	Massed versus embedded instruction may differ across participants, but due to no intraparticipant replication with counterbalancing, differences may also be due to differential behavior difficulty.
Trial Ordering	Interspersing known (mastered) trials may result in slower learning.

Figure 1. PRISMA diagram outlining search and screening procedures. To be included in the descriptive synthesis, comparisons were required to include a baseline condition. To be included in the outcomes analysis, studies had to include acceptable reliability data and *either* (a) a control set or (b) time-lagged comparison. Information about specific designs that were included and excluded for each source are available in the Supplemental Coding File.

References

One asterisk means a study was included for descriptive coding only. Two asterisks mean the study was included in descriptive and outcomes coding. Note that this indicates at least one comparison in the source was coded for outcomes, not necessarily that all comparisons were.

- *Ackerlund Brandt, J. A., Weinkauff, S., Zeug, N., & Klatt, K. P. (2016). An evaluation of constant time delay and simultaneous prompting procedures in skill acquisition for young children with Autism. *Education and Training in Autism and Developmental Disabilities, 51*, 55-66.
- *Addison, L. (2006). An examination of the differential outcomes effect when teaching discriminations to children with autism and other developmental disabilities (Doctoral dissertation, Louisiana State University).
- *Akmanoglu, N., Kurt, O., & Kapan, A. (2015). Comparison of simultaneous prompting and constant time delay procedures in teaching children with autism the responses to questions about personal information. *Educational Sciences: Theory and Practice, 15*, 723-737.
- **Apple, A. L. (2005). A comparison of the use of constant time delay alone and constant time delay with instructive feedback to teach children with autism to discriminate stimuli by function, feature, and class. ProQuest Dissertations & Theses Global. (305421964).
- **Ault, M. J., Gast, D. L., & Wolery, M. (1988). Comparison of progressive and constant time-delay procedures in teaching community-sign word reading. *American Journal on Mental Retardation, 92*, 44-56.
- *Ault, M. J., Wolery, M., Gast, D. L., Doyle, P. M., & Martin, C. P. (1990). Comparison of predictable and unpredictable trial sequences during small-group instruction. *Learning Disability Quarterly, 13*, 12-29.
- **Bennett, D. L., Gast, D. L., Wolery, M., & Schuster, J. (1986). Time delay and system of least prompts: A comparison in teaching manual sign production. *Education and Training of the Mentally Retarded, 21*, 117-129.
- *Bertsch, K. M. (2002). A comparison of one-to-one and small group instruction for young children with autism: focus on effective teaching and behavior management. ProQuest Dissertations & Theses Global. (305501301).
- **Boulware, G. (2001). A comparison of simultaneous and most -to -least prompting procedures in teaching receptive identification of pictures to toddlers with autism. ProQuest Dissertations & Theses Global. (250778791).
- **Boutain, A. R. (2012). A comparison of graduated guidance and simultaneous prompting in teaching receptive language skills to children with autism. ProQuest Dissertations & Theses Global. (1034439160).
- **Call, N. (2013). The effects of intertrial intervals on receptive tasks for young children with autism (Doctoral dissertation, University of Kansas).
- **Carroll, R. A., Joachim, B. T., St Peter, C. C., & Robinson, N. (2015). A comparison of error-correction procedures on skill acquisition during discrete-trial instruction. *Journal of Applied Behavior Analysis, 48*, 257-273. doi: 10.1002/jaba.205
- *Carroll, R. A., Kodak, T., & Adolf, K. J. (2016). Effect of delayed reinforcement on skill acquisition during discrete-trial instruction: Implications for treatment-integrity errors in

- academic settings. *Journal of Applied Behavior Analysis*, 49, 176-181. doi: 10.1002/jaba.268
- **Carroll, R. A., Kodak, T., & Fisher, W. W. (2013). An evaluation of programmed treatment-integrity errors during discrete-trial instruction. *Journal of Applied Behavior Analysis*, 46, 379-394. doi: 10.1002/jaba.49
- *Casey, S. D. (2008). A comparison of within-and across-session progressive time delay procedures for teaching sight words to individuals with cognitive delays. *The Behavior Analyst Today*, 9, 162.
- *Celik, S., & Vuran, S. (2014). Comparison of direct instruction and simultaneous prompting procedure on teaching concepts to individuals with intellectual disability. *Education and Training in Autism and Developmental Disabilities*, 49, 127-144.
- **Cengher, M., Shamoun, K., Moss, P., Roll, D., Feliciano, G., & Fienup, D. M. (2016). A comparison of the effects of two prompt-fading strategies on skill acquisition in children with autism spectrum disorders. *Behavior Analysis in Practice*, 9, 115-125. doi: 10.1007/s40617-015-0096-6
- *Cividini-Motta, C. (2014). Teaching verbal behavior: A systematic comparison of commonly used echoic teaching procedures (Doctoral dissertation, Western New England University).
- *Clark, C. P. (2000). The differential effects of simplified speech versus typical speech on response accuracy in discrete trial teaching formats with prekindergarten age students with autism (Doctoral dissertation, Florida Atlantic University).
- **Clements, A. (2014). An evaluation of the effect of attention during descriptive praise on skill acquisition during discrete trial instruction (Master's Thesis, University of Nebraska at Omaha).
- *Collins, B. C., Evans, A., Creech-Galloway, C., Karl, J., & Miller, A. (2007). Comparison of the acquisition and maintenance of teaching functional and core content sight words in special and general education settings. *Focus on Autism and Other Developmental Disabilities*, 22, 220-233. doi: 10.1177/10883576070220040401
- **Colozzi, G. A., Ward, L. W., & Crotty, K. E. (2008). Comparison of simultaneous prompting procedure in 1: 1 and small group instruction to teach play skills to preschool students with pervasive developmental disorder and developmental disabilities. *Education and Training in Developmental Disabilities*, 43, 226-248.
- *Davis, M. (2011). Comparison of massed trial instruction vs. activity-based instruction using simultaneous prompting with young children who are at risk. (Doctoral Dissertation, University of Kentucky).
- *Day, H. (1987). Comparison of two prompting procedures to facilitate skill acquisition among mentally retarded adolescents. *American Journal of Mental Deficiency*, 92, 366-372.
- *Dittlinger, L. H., & Lerman, D. C. (2011). Further analysis of picture interference when teaching word recognition to children with autism. *Journal of Applied Behavior Analysis*, 44, 341-349. doi: 10.1901/jaba.2011.44-341
- *Dufour, M. M., & Lanovaz, M. J. (2017). Comparing two methods to promote generalization of receptive identification in children with autism spectrum disorders. *Developmental Neurorehabilitation*, 20, 463-474.
- *Fentress, G. M., & Lerman, D. C. (2012). A comparison of two prompting procedures for teaching basic skills to children with autism. *Research in Autism Spectrum Disorders*, 6, 1083-1090. doi: 10.1016/j.rasd.2012.02.006

- *Ferkis, M. A., Belfiore, P. J., & Skinner, C. H. (1997). The effects of response repetitions on sight word acquisition for students with mild disabilities. *Journal of Behavioral Education*, 7, 307-324.
- *Fossett, B., & Mirenda, P. (2006). Sight word reading in children with developmental disabilities: A comparison of paired associate and picture-to-text matching instruction. *Research in Developmental Disabilities*, 27, 411-429. doi: 10.1016/j.ridd.2005.05.006
- *Gard, M. (2012). An evaluation of sight word instruction procedures for children with autism: Interspersed trial training (Doctoral dissertation, University of Kansas).
- *Geiger, K. B., Carr, J. E., LeBlanc, L. A., Hanney, N. M., Polick, A. S., & Heinicke, M. R. (2012). Teaching receptive discriminations to children with autism: a comparison of traditional and embedded discrete trial teaching. *Behavior Analysis in Practice*, 5, 49-59. doi: 10.1007/BF03391823
- **George, P. R. (1997). A comparison of group and individual written attentional responses to enhance targeted, observational and incidental learning of social studies facts in a middle school, inclusion setting. ProQuest Dissertations & Theses Global. (304343500).
- **Godby, S., Gast, D. L., & Wolery, M. (1987). A comparison of time delay and system of least prompts in teaching object identification. *Research in Developmental Disabilities*, 8, 283-305. doi: 10.1016/0891-4222(87)90009-6
- **Gorgan, E. M. (2017). Assessing treatment options for pre-existing prompt dependence in children with developmental disabilities (Doctoral dissertation, University of Wisconsin – Milwaukee).
- *Griffen, A. K., Schuster, J. W., & Morse, T. E. (1998). The acquisition of instructive feedback: A comparison of continuous versus intermittent presentation schedules. *Education and Training in Mental Retardation and Developmental Disabilities*, 33, 42-61.
- *Grow, L. L., Kodak, T., & Clements, A. (2017). An evaluation of instructive feedback to teach play behavior to a child with autism spectrum disorder. *Behavior Analysis in Practice*, 10, 313-317.
- **Harrell, P. L. (1990). A comparison of the efficiency of interdependent and independent group contingencies with mildly handicapped elementary-aged students concepts using time delay instructional procedures (Doctoral dissertation, University of Kentucky).
- **Hawkins, S. R. (2008). Embedding constant time delay instruction to teach discrete tasks to preschool children who have significant disabilities in a one to one and small group format: a comparison study (Doctoral dissertation, University of Kentucky).
- **Head, K. D., Collins, B. C., Schuster, J. W., & Ault, M. J. (2011). A comparison of simultaneous prompting and constant time delay procedures in teaching state capitals. *Journal of Behavioral Education*, 20, 182-202. doi: 10.1007/s10864-011-9127-8
- *Henrickson, M. L., Rapp, J. T., & Ashbeck, H. A. (2015). Teaching with massed versus interspersed trials: Effects on acquisition, maintenance, and problem behavior. *Behavioral Interventions*, 30, 36-50. doi: 10.1002/bin.1396
- *Holcombe, A., Wolery, M., & Snyder, E. (1994). Effects of two levels of procedural fidelity of constant time delay on children's learning. *Journal of Behavioral Education*, 4, 49-73. doi: 10.1007/BF01560509
- **Humphreys, T., Polick, A. S., Howk, L. L., Thaxton, J. R., & Ivancic, A. P. (2013). An evaluation of repeating the discriminative stimulus when using least-to-most prompting to teach intraverbal behavior to children with autism. *Journal of Applied Behavior Analysis*, 46, 534-538. doi: 10.1002/jaba.43

- *Ingvarsson, E. T., & Hollobaugh, T. (2011). A comparison of prompting tactics to establish intraverbals in children with autism. *Journal of Applied Behavior Analysis*, 44, 659-664.
- *Ingvarsson, E., & Le, D. (2011). Further evaluation of prompting tactics for establishing intra-verbal responding in children with autism. *The Analysis of Verbal Behavior*, 27, 75-93.
- *Iacono, T. A. (1990). A comparison of unimodal and multimodal augmentative communication language intervention techniques for children with intellectual disabilities (Doctoral dissertation, University of Nebraska).
- *Jameson, J. M., McDonnell, J., Johnson, J. W., Riesen, T. J., & Polychronis, S. (2007). A comparison of one-to-one embedded instruction in the general education classroom and one-to-one massed practice instruction in the special education classroom. *Education and Treatment of Children*, 30, 23-44.
- *Joachim, B. T., & Carroll, R. A. (2018). A comparison of consequences for correct responses during discrete-trial instruction. *Learning and Motivation*, 62, 15-28.
- *Johnson, P., Schuster, J., & Bell, J. K. (1996). Comparison of simultaneous prompting with and without error correction in teaching science vocabulary words to high school students with mild disabilities. *Journal of Behavioral Education*, 6, 437-458. doi: 10.1007/BF02110516
- **Julien, H. M., & Reichle, J. (2016). A comparison of high and low dosages of a component of milieu teaching strategies for two preschool-age learners with autism spectrum disorder. *Language, Speech, and Hearing Services in Schools*, 47, 87-98. doi: 10.1044/2015_LSHSS-15-0035
- *Kodak, T., Fuchtman, R., & Paden, A. (2012). A comparison of intraverbal training procedures for children with autism. *Journal of Applied Behavior Analysis*, 45, 155-160.
- **Leaf, J. B., Alcalay, A., Leaf, J., Tsuji, K., Kassardjian, A., Dale, S., ... & Leaf, R. (2016). Comparison of most-to-least to error correction for teaching receptive labelling for two children diagnosed with autism. *Journal of Research in Special Educational Needs*, 16, 217-225. doi: 10.1111/1471-3802.12067
- **Leaf, J. B., Cihon, J. H., Ferguson, J. L., McEachin, J., Leaf, R., & Taubman, M. (2018). Evaluating three methods of stimulus rotation when teaching receptive labels. *Behavior Analysis in Practice*, 11, 1-16. doi: 10.1007/s40617-018-0249-5
- **Leaf, J. B., Leaf, J. A., Alcalay, A., Kassardjian, A., Tsuji, K., Dale, S., ... & Leaf, R. (2016). Comparison of most-to-least prompting to flexible prompt fading for children with autism spectrum disorder. *Exceptionality*, 24, 109-122. doi: 10.1080/09362835.2015.1064419
- **Leaf, J. B., Leaf, R., Taubman, M., McEachin, J., & Delmolino, L. (2014). Comparison of flexible prompt fading to error correction for children with autism spectrum disorder. *Journal of Developmental and Physical Disabilities*, 26, 203-224. doi: 10.1007/s10882-013-9354-0
- **Leaf, J. B., Sheldon, J. B., & Sherman, J. A. (2010). Comparison of simultaneous prompting and no-no prompting in two-choice discrimination learning with children with autism. *Journal of Applied Behavior Analysis*, 43, 215-228. doi: 10.1901/jaba.2010.43-215
- **Leaf, J. B., Townley-Cochran, D., Mitchell, E., Milne, C., Alcalay, A., Leaf, J., ... & Oppenheim-Leaf, M. L. (2016). Evaluation of multiple-alternative prompts during tact training. *Journal of Applied Behavior Analysis*, 49, 399-404. doi: 10.1002/jaba.289
- **Leaf, J. B., Tsuji, K. H., Lentell, A. E., Dale, S. E., Kassardjian, A., Taubman, M., ... & Oppenheim-Leaf, M. L. (2013). A comparison of discrete trial teaching implemented in a

- one-to-one instructional format and in a group instructional format. *Behavioral Interventions*, 28, 82-106. doi: 10.1002/bin.1357
- *Ledford, J. R., Barton, E., Hardy, J., Elam, K., Seabolt, J., Shanks, M., ... & Kaiser, A. (2016). What equivocal data from single case comparison studies reveal about evidence-based practices in early childhood special education. *Journal of Early Intervention*, 38, 79-91. doi: 10.1177/1053815116648000
- **Ledford, J. R., Chazin, K. T., Harbin, E. R., & Ward, S. E. (2017). Massed trials versus trials embedded into game play: Child outcomes and preference. *Topics in Early Childhood Special Education*, 37, 107-120. doi: 10.1177/0271121417704703
- **Majdalany, L., Wilder, D., Greif, A., Mathisen, D., & Saini, V. (2014). Comparing massed-trial instruction, distributed-trial instruction, and task interspersal to teach tacts to children with autism spectrum disorders. *Journal of Applied Behavior Analysis*, 47, 657-662. doi: 10.1002/jaba.149
- **Majdalany, L., Wilder, D. A., Smeltz, L., & Lipschultz, J. (2016). The effect of brief delays to reinforcement on the acquisition of tacts in children with autism. *Journal of Applied Behavior Analysis*, 49, 411-415. doi: 10.1002/jaba.282
- *McDonnell, J., Johnson, J. W., Polychronis, S., Riesen, T., Kercher, K., & Jameson, M. (2006). Comparison of one-to-one embedded instruction in general education classes with small group instruction in special education classes. *Education and Training in Developmental Disabilities*, 41, 125-138.
- *McGhan, A., & Lerman, D. C. (2013). Assessment of error-correction procedures for learners with autism. *Journal of Applied Behavior Analysis*, 46, 626-639. doi: 10.1002/jaba.65
- *Nopprapun, M., & Holloway, J. (2014). A comparison of fluency training and discrete trial instruction to teach letter sounds to children with ASD: Acquisition and learning outcomes. *Research in Autism Spectrum Disorders*, 8, 788-802.
- *O'Neill, S.J., McDowell, C., & Leslie, C.J. (2018). A comparison of prompt delays with trial-and-error instruction in conditional discrimination training. *Behavior Analysis in Practice*. doi: 10.1007/s40617-018-0261-9
- *Odluyurt, S., Tekin-Iftar, E., & Adalioglu, I. (2012). Does treatment integrity matter in promoting learning among children with developmental disabilities? *Topics in Early Childhood Special Education*, 32, 143-150. doi: 10.1177/0271121410394208
- **Owisany, J.M. (2017). An evaluation of the effects of treatment integrity errors on skill acquisition during discrete-trial instruction (Master's thesis, University of West Virginia University)
- *Parker, E.D. (2009). Tact repertoires and measures of efficiency: comparing the effects of two behavioral intervention models with students with developmental disabilities (Doctoral dissertation, The Ohio State University).
- *Pesantez, R. (2012). Comparing prompt delay and total communication for training vocal intraverbals in children with autism. Dissertations & Theses Global. (1251630647).
- **Polick, A.S. (2011). An investigation of general and descriptive praise in teaching language to children with autism (Doctoral dissertation, Auburn University).
- *Polychronis, S. C., McDonnell, J., Johnson, J. W., Riesen, T., & Jameson, M. (2004). A comparison of two trial distribution schedules in embedded instruction. *Focus on Autism and Other Developmental Disabilities*, 19, 140-151. doi: 10.1177/10883576040190030201

- **Pufpaff, L. A., Blischak, D. M., & Lloyd, L. L. (2000). Effects of modified orthography on the identification of printed words. *American Journal on Mental Retardation*, 105, 14-24.
- *Ramirez, S.R. (2017). Comparing the efficacy of picture card stimuli to storybook stimuli in teaching morphologic structures to children with expressive language disorders (Master's thesis, California State University).
- **Reichow, B., & Wolery, M. (2011). Comparison of progressive prompt delay with and without instructive feedback. *Journal of Applied Behavior Analysis*, 44, 327-340. doi: 10.1901/jaba.2011.44-327
- *Riesen, T., McDonnell, J., Johnson, J. W., Polychronis, S., & Jameson, M. (2003). A comparison of constant time delay and simultaneous prompting within embedded instruction in general education classes with students with moderate to severe disabilities. *Journal of Behavioral Education*, 12, 241-259. doi: 10.1023/A:1026076406656
- *Roest, E.R. (2008). A comparative study of the effects of fluency training and discrete trial training on the acquisition, stimulus generalization and retention of noun labels in children with autism (Doctoral dissertation, University of Connecticut).
- *Schlosser, R.W., Sigafoos, J., Luiselli, J.K., Angermeier, K., Harasymowycz, U., Schooley, K., & Belfiore, P.J. (2007). Effects of synthetic speech output on requesting and natural speech production in children with autism: A preliminary study. *Research in Autism Spectrum Disorders*, 1, 139-163. doi:10.1016/j.rasd.2006.
- **Schuster, J. W., Griffen, A. K., & Wolery, M. (1992). Comparison of simultaneous prompting and constant time delay procedures in teaching sight words to elementary students with moderate mental retardation. *Journal of Behavioral Education*, 2, 305-325. doi: 10.1007/BF00948820
- **Shook, S. L. (1999). Teaching children with autism to ask questions in integrated preschool settings: a comparison of constant and progressive time delay (Doctoral dissertation, University of Washington).
- **Singleton, D. K., Schuster, J. W., Morse, T. E., & Collins, B. C. (1999). A comparison of antecedent prompt and test and simultaneous prompting procedures in teaching grocery words to adolescents with mental retardation. *Education and Training in Mental Retardation and Developmental Disabilities*, 34, 182-199.
- *Smith, S. L. (2016). Comparing error correction procedures: Incorporating effective and efficient teaching with learner preference (Doctoral dissertation, Temple University).
- *Soluaga, D., Leaf, J. B., Taubman, M., McEachin, J., & Leaf, R. (2008). A comparison of flexible prompt fading and constant time delay for five children with autism. *Research in Autism Spectrum Disorders*, 2, 753-765. doi: 10.1016/j.rasd.2008.03.005
- **Spino, M. A. (2013). The effects of two schedules of instruction with constant time delay on the receptive word learning skills of preschool children with developmental delays. (Doctoral dissertation, The University of Toledo).
- *Swain, R., Lane, J. D., & Gast, D. L. (2015). Comparison of constant time delay and simultaneous prompting procedures: Teaching functional sight words to students with intellectual disabilities and autism spectrum disorder. *Journal of Behavioral Education*, 24, 210-229. doi: 10.1007/s10864-014-9209-5
- **Tekin, E., & Kircaali-Iftar, G. (2002). Comparison of the effectiveness and efficiency of two response prompting procedures delivered by sibling tutors. *Education and Training in Mental Retardation and Developmental Disabilities*, 37, 283-299.

- **Toussaint, K. A., Kodak, T., & Vladescu, J. C. (2016). An evaluation of choice on instructional efficacy and individual preferences among children with autism. *Journal of Applied Behavior Analysis*, 49, 170-175. doi: 10.1002/jaba.263
- *Valentino, A. L., Shillingsburg, M. A., & Call, N. A. (2012). Comparing the effects of echoic prompts and echoic prompts plus modeled prompts on intraverbal behavior. *Journal of Applied Behavior Analysis*, 45, 431-435. doi: 10.1901/jaba.2012.45-431
- *Vedora, J., Meunier, L., & Mackay, H. (2009). Teaching intraverbal behavior to children with autism: A comparison of textual and echoic prompts. *The Analysis of Verbal Behavior*, 25, 79-86. doi: 10.1007/BF03393072
- *Vedora, J., & Conant, E. (2015). A comparison of prompting tactics for teaching intraverbals to young adults with autism. *The Analysis of Verbal Behavior*, 31, 267-276. doi: 10.1007/s40616-015-0030-6
- *Vladescu, J. C., & Kodak, T. M. (2013). Increasing instructional efficiency by presenting additional stimuli in learning trials for children with autism spectrum disorders. *Journal of Applied Behavior Analysis*, 46, 805-816. doi: 10.1002/jaba.70
- *Volkert, V., Lerman, D., Trosclair, N., Addison, L., & Kodak, T. (2008). An exploratory analysis of task-interspersal procedures for teaching object labels to children with autism. *Journal of Applied Behavior Analysis*, 41, 335-350. doi: 10.1901/jaba.2008.41-335
- *Wallace, A. M., Bechtel, D. R., Heatter, S., & Barry, L. M. (2016). A comparison of prompting strategies to teach intraverbals to an adolescent with down syndrome. *The Analysis of Verbal Behavior*, 32, 225-232. doi: 10.1007/s40616-016-0058-2
- **Waugh, R. E., Alberto, P., & Fredrick, L. (2011). Effects of error correction during assessment probes on the acquisition of sight words for students with moderate intellectual disabilities. *Research in Developmental Disabilities*, 32, 47-57. doi: 10.1016/j.ridd.2010.08.007