

A Meta-Analysis of the Graduated Guidance Procedure

The Journal of Special Education
1–12

© Hammill Institute on Disabilities 2023

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/00224669231200461

journalofspecialeducation.sagepub.com



Elif Tekin-Iftar, PhD¹ , Melinda Jones Ault, PhD²,
Belva C. Collins, EdD³, Seray Olcay, PhD⁴,
H. Deniz Degirmenci, PhD¹, and Orhan Aydın, PhD^{1,5}

Abstract

We conducted a descriptive analysis and meta-analysis of single-case research design (SCRD) studies investigating the effectiveness of the graduated guidance procedure. Once we identified studies through electronic databases and reference lists, we used What Works Clearinghouse (WWC) Standards to evaluate each study. Then, we described studies in terms of various descriptive variables, calculated effect sizes through three non-parametric effect size methods, and analyzed results across studies. Results showed 11 of the 27 studies met WWC Standards or met standards with reservation. Of the 11, seven studies resulted in a large effect. We found the graduated guidance procedure to be an evidence-based practice when evaluating the findings against contemporary evidence standards. However, this review also showed that the majority of the reviewed studies ($n = 20$) had no effects and only one third of the studies had moderate or strong effects. Implications for researchers and practitioners are discussed.

Keywords

graduated guidance, response prompting, systematic instruction

When possible, instructors of students with disabilities should use evidence- or research-based practices during instruction (Cook & Cook, 2011). This allows them to select strategies they can be reasonably assured, based on high-quality experimental studies, will be effective in teaching or remediating a particular behavior with a particular population of learners (Ledford et al., 2019). This is especially important, given many students may require more intensive instruction in higher dosage to acquire content knowledge and become proficient in skills than students without disabilities. The more efficiently a student can be taught, the more knowledge and skills the student can be expected to learn in time allotted for instruction. Just as it is necessary to prioritize content and skills for instruction when developing a student's individualized education program, it is also important to prioritize instructional strategies in terms of what can be expected to be most effective and efficient for teaching specific content, whether academic or adaptive. This is why researchers have identified high-leverage practices (McLeskey et al., 2017) to include explicit and intensive systematic instruction with scaffolding when teaching students with disabilities.

Response prompting procedures (Collins, 2022) are a specific category of systematic instruction used to teach new behaviors to learners in which the instructor provides a stimulus (e.g., task direction) followed by a controlling

prompt to assist the student in making a correct response. Prompts (e.g., models, physical guidance) are then systematically faded over time in numerous ways as no longer needed. This may include decreasing intensity or intrusiveness of the prompt or increasing the interval in which the student can respond before a prompt is delivered. Prompts should be individualized according to type of task being taught, type of disability characteristics of student, and type of student support required.

Single-case research designs (SCRDs) are typically used to investigate effectiveness and efficiency of response prompting procedures (Ledford & Severini, 2019). These studies focus on teaching a specific skill type (e.g., sight word vocabulary, life skill) to a small number of students with a specific disability type (e.g., autism spectrum disorder [ASD], intellectual disability [ID]) across a specific age

¹Anadolu University, Eskisehir, Turkey

²University of Kentucky, Lexington, USA

³University of North Carolina-Charlotte, USA

⁴Hacettepe University, Ankara, Turkey

⁵Erzincan Binali Yildirim University, Turkey

Corresponding Author:

Elif Tekin-Iftar, Anadolu Universitesi, Engelliler Arastirma Enstitusu,
26470 Eskisehir, Turkey.

Email: eltekin@anadolu.edu.tr

range (e.g., preschool, elementary, secondary). For a response prompting procedure to be considered evidence-based, it is necessary to build a sufficient research base that meets stringent requirements set forth by researchers (e.g., Kratochwill et al., 2013; What Works Clearinghouse [WWC], 2022). Procedures researchers have determined to be evidence-based through meta-analyses include time delay (e.g., Browder et al., 2009), simultaneous prompting (Tekin-Iftar et al., 2019), and system of least prompts (Shepley et al., 2018). The research base includes a sufficient number of rigorous studies with SCRDs conducted by researchers from different geographical areas across participants concluding the procedures likely to be effective in teaching specific skills to students with specific disabilities, if used with fidelity.

The graduated guidance (GG) response prompting procedure has been in the literature since first described by Azrin and Foxx in 1971 (Collins, 2022; Wolery et al., 1992). In this procedure, the instructor typically uses a physical prompt to guide a student in making a correct response, making moment-by-moment decisions on level of physical prompting needed to ensure the response, until the student can perform the desired behavior. That is, as the student is responding, the instructor makes decisions to provide or withhold physical prompting needed in the moment to maintain correct responding. The procedure is so natural that those who use it may not recognize they are using a response prompting procedure. For example, instructors, including parents or caregivers, may use GG when teaching a child to perform fine motor (e.g., feeding or dressing oneself) or gross motor (e.g., riding a bicycle) skills. The instructor simply shadows the student's movement (i.e., keeps hands close by student without touching or influencing movements) providing physical guidance as needed at any given moment to keep the student responding correctly (Collins, 2022).

The research literature contains examples of the GG procedure used as a single strategy or combined with another strategy (e.g., GG plus video modeling, time delay, most-to-least). GG can be used with the time delay procedure by delivering a physical prompt after a designated delay or response interval for performing a skill has elapsed (McLay et al., 2017; Sigafoos et al., 2018). GG can be used with the most-to-least prompting procedure by systematically fading a physical prompt by intensity (e.g., hand over hand to partial physical to gentle nudge) or location (e.g., hand on hand to wrist to elbow; Jimenez & Alamer, 2018). In this case, the procedure would be classified as most-to-least rather than GG because criterion levels were used before moving to a less intrusive prompt level. As an example of a variation of the GG procedure, Akmanoglu et al. (2014) provided video modeling followed by the GG procedure in teaching role playing skills to children with ASD, pairing GG with a verbal prompt while fading from full physical guidance to partial physical to shadowing.

Although the literature contains variations of the GG procedure, we focused the current meta-analysis on studies that used the basic form of the GG procedure as a single strategy (e.g., Cattik & Odluyurt, 2017; Horsman, 2018). That is, we reviewed studies in which the instructional procedures consisted of the instructor shadowing the response of a student while making moment-by-moment decisions on providing or reducing the physical prompt to facilitate a correct response by the student. The instructor provides the least intrusive physical prompt needed to ensure a correct response. In this case, the prompt naturally fades when no longer needed. By focusing on this narrow definition in the meta-analysis, we could determine whether GG was effective without having to consider other confounding variables that might have made it more or less effective. Thus, the research question was as follows:

Research Question 1 (RQ1): Is the GG procedure an evidence-based practice for teaching skills to students with disabilities?

Method

Search Procedures

We systematically searched the literature to locate studies investigating the GG procedure until July 2020 without specifying a starting date so as to include earlier studies. We searched databases that included Web of Science, JSTOR, Worldcat.org, PsycINFO, PsycARTICLES, EBSCOhost, ScienceDirect, and ERIC. To ensure a comprehensive analysis, we also searched databases to locate GG studies in the gray literature: ProQuest, UMI, Thesis Global, and Open Dissertation. We used various combinations of relevant keywords (e.g., *graduated guidance*, *physical prompting*, *manual prompting*, *flexible prompting*), adding “or” between keywords. Combinations of keywords searched can be obtained from the authors upon request. Three researchers independently conducted searches, then shared obtained studies electronically by placing them in a shared, single folder. The folder included 529 articles and 53 unpublished theses ($n = 11$) or dissertations ($n = 42$).

We reviewed titles and abstracts of obtained journal articles and designated 219 articles to be reviewed for further analysis. We excluded duplications (i.e., same article located by at least two researchers), leaving 169 journal articles to be reviewed. After removing unpublished thesis or dissertation studies where we could not access the full text, or that were duplicated, we retained 34 unpublished studies to be reviewed. This left a total of 203 studies. Two researchers independently coded each of these studies according to research methodologies under three categories: (a) SCRDs, (b) review studies, and (c) other research designs. This left 167 (92%) studies where investigators

used SCRDs. Regarding the designs of the studies, the inter-coder agreement between the two researchers was 92% and 91.6% for included and excluded studies, respectively (inclusion and exclusion procedures explained below). The two researchers reviewed reference lists of these journal articles and unpublished studies, resulting in 2,768 additional studies. They conducted an ancestral search by reviewing titles of additional studies from reference lists of these studies. If the title appeared to be related to the present study, they reviewed its abstract. If the abstract appeared to be related, they analyzed the full text. As a result, they identified 10 additional articles for a total of 177 studies to be included in the analysis for the present study.

Inclusion and Exclusion Criteria

We used inclusion criteria for studies in the analysis as follows: (a) published in English in internationally disseminated peer-reviewed journals, (b) unpublished theses or dissertations in English, (c) GG procedure used as the independent variable in teaching skills in a demonstration study, (d) GG used as one independent variable in a comparison study, (e) GG described in detailed and replicable manner (e.g., levels of physical prompts, moment-by-moment decision making), and (f) study conducted using SCRD.

We excluded studies in which (a) the GG procedure was not the independent variable systematically manipulated, (b) at least one participant did not have a disability, (c) appropriate data were not provided for visual/graphic analysis, and (d) GG was combined with another strategy. Supplemental Figure 1 shows the process and number of studies identified for inclusion and analysis. Two researchers independently coded the studies considering inclusion/exclusion criteria. Both included 37 studies for further analyses and excluded 68 studies. Agreement between the two researchers was 59% across 177 studies. They did not agree on 72 studies. Possible reasons for obtaining low agreement included disagreements on (a) if GG procedure was the sole independent variable, (b) if the prompt fading strategy was based on implementer's moment-by-moment decision-making, and (c) if GG procedure was combined with another response prompting procedure (e.g., time delay).

The two researchers discussed each indicator to re-code the 72 studies on which they failed to agree on inclusion or exclusion. This resulted in 100% agreement to exclude 47 studies and failure to reach consensus on 25 studies, resulting in 65% inter-rater agreement. Then they discussed the studies on which they disagreed with the team's senior researchers (two in the United States, one in Turkey.) The senior researchers individually coded the studies and discussed until they reached 100% consensus resulting in eight studies included for further analyses. Also, based on discussion of the three researchers, the senior researcher in Turkey excluded 18 studies that had been included earlier and

shared these with the other senior researchers for their evaluation. They also reached 100% consensus on these studies. As a result of these discussions, 27 studies (19 journal articles, eight unpublished thesis/dissertation studies) met criteria for inclusion in the review. Once we identified included studies, one researcher conducted a forward chaining search on Google Scholar to locate articles cited in the included study. From this search, the researcher located 1,779 citations. If titles of these studies appeared related to the GG procedure, the researcher read abstracts. As a result, the researcher located five studies for further analysis. Two researchers independently coded and decided to exclude these five studies with 100% agreement.

Procedures for Evaluating Quality Indicators of the Studies

We used the research design standards recommended by Kratochwill et al. (2013), the contemporary standards at the time of the review, to evaluate the design quality of the 27 identified studies. We modified and added two indicators to Kratochwill et al.'s standards about the treatment integrity of the studies (items d and e that follow). We created a data sheet to determine presence and absence of each indicator within eight categories: (a) systematic manipulation of independent variable, (b) collection of interobserver data for at least 20% of all sessions, (c) interobserver agreement of at least 80% of all sessions, (d) collection of treatment integrity data for at least 20% of all sessions, (e) treatment integrity of at least 80% of all sessions, (f) at least three demonstrations of effect, (g) at least five data points per condition (to meet standards), (h) at least three data points per condition (to meet standards with reservation), (i) classification of design standards, and (j) classification of evidence for effectiveness. (Note we did not take items d and e into consideration while evaluating classification of design standards.) Prior to evaluating the articles, three researchers discussed and listed decision rules for each indicator. One researcher trained coders on rules. Then, they independently coded three randomly selected studies and reached 100% consensus through discussion. After that, two researchers coded indicators of each article (analyses shown in Supplemental Table 1). The researchers examined each tier in a study to determine the presence of an indicator, coding "yes" if all tiers met the indicator in the study. Even if a study failed to meet an indicator in a single tier, the researchers coded that indicator as "no" for the study. Classification of standards (second to last column in Supplemental Table 1) were coded according to definitions that follow.

Even if a study failed to meet an indicator in a single tier, we coded that indicator as "no" for the study. Studies which *met criteria* (stated above) *from a to g* were coded as *meets standards (MS)*, studies that *met criteria from a to h except*

g were coded as *MS with reservations (MS-R)*, studies that met criterion e but did not meet at least one of the criteria between a and f were coded as *does not meet standards (nMS)* in Supplemental Table 1. Classification of evidence of effectiveness (item i in preceding paragraph and last column in Supplemental Table 1) were coded using visual analysis of studies categorized as *MS* and *MS-R*. We retained studies that met design standards in all tiers for visual and descriptive analyses. To assess effects, we considered six outcome measures (i.e., level, trend, variability, immediacy of effects, overlap, consistency of data patterns across similar phases) within and between conditions (Kratochwill et al., 2013). If a study provided demonstration of an effect in all outcome measures, we categorized it as *strong evidence*. If a study provided three demonstrations of an effect and also included at least one demonstration of non-effect, we categorized it as *moderate evidence*. If a study did not provide at least three, temporally distinct, demonstrations of an effect, we categorized it as *no evidence*. Demonstration of an effect through visual analysis in all outcomes for studies is shown in Supplemental Table 2. We retained studies categorized as having either *strong* or *moderate* evidence for calculation of effect size estimations.

Procedures for Conducting Descriptive Analysis of the GG Studies

For each study coded as *MS* and *MS-R* as recommended by Kratochwill et al. (2013), the two researchers who coded design standards also coded the following data for descriptive analysis: (a) characteristics of participants; (b) skill area taught; (c) setting and teaching format; (d) design; (e) intervention; (f) social validity, maintenance, and generalization; (g) intensity and duration of intervention and (h) overall outcomes. Supplemental Table 3 displays compiled data.

Intervention Effect Calculations

There is a debate about which method is most appropriate in synthesizing SCRDs. While analyzing effect size in SCRDs, contradictory findings can be obtained (Chen et al., 2019); thus, using more than one effect size analysis method has been suggested (Kratochwill et al., 2013). For this reason, we used three effect size analysis methods: percentage of non-overlapping data (PND), Tau-*U*, and improvement rate difference (IRD). The most frequently used method to calculate effect size is PND (Scruggs & Mastropieri, 2001), which has advantages over other methods for detecting single-case effect size estimates: It does not require linearity, it is easy to calculate, and significant correlations can be found between PND and other effect size estimates (Campbell, 2004). However, a noted limitation of PND is that standard errors and confidence intervals cannot be calculated (Parker et al., 2007). Tau-*U* is another non-overlap

method suitable for any type of distribution and scale used to calculate effect size; it controls for undesirable positive baseline trend in studies (Parker et al., 2011). Although data dependence (i.e., autocorrelation) may visibly distort Tau-*U* values (Parker et al., 2011; Solomon et al., 2015), it gives more robust results than other nonoverlap effect size estimates even if serial dependency is observed between single-case data (Tarlow, 2017). In a recent study, Barnard-Brak et al. (2021) showed that Tau-*U* calculations are not significantly affected by the degree of autocorrelation. However, Tekin-Iftar et al. (2019) indicated a high correlation between PND, which is significantly affected by serial dependency (Barnard-Brak et al., 2021), and Tau-*U* analyses in their study. The third effect size method, IRD, expresses the difference in successful performance between baseline and intervention conditions. It detects and controls possible increase in baseline condition and detects increase in intervention condition and analyzes the difference (Parker et al., 2009). We conducted baseline-intervention comparison in all effect size estimates in the reviewed studies. We calculated PND by hand, identifying the highest data point in baseline condition and then identifying intervention data points exceeding that point. Then, we calculated PND by dividing total number of intervention data points above the highest data point by total number of data points of comparison condition and multiplying by 100. We calculated Tau-*U* and IRD scores using the web-based Tau-*U* calculator at <http://www.singlecaseresearch.org> (Vannest et al., 2016). Since confidence intervals for Tau-*U* values are obtained in the web-based computation engine, we also report 95% confidence intervals (CI₉₅) for them. This also enables us to generate forest plots for Tau-*U* calculations.

Based on guidelines, we considered PND scores at or above 90% as *very effective*, between 70% and 90% as *effective*, between 50% and 70% as *questionable*, and below 50% as *ineffective* (Scruggs & Mastropieri, 2001). We considered Tau-*U* scores at or above 93% as *strong effect*, between 66% and 92% as *medium to high effect*, and between 0% and 65% as *small effect* (Parker & Vannest, 2009). Parker et al. (2009) indicated IRD scores at or below 50% as *small effect*, between 50% and 70% as *moderate effect*, and 70% or above as *large effect*. In the present review, effect sizes between the included studies were calculated in a random effect manner because the studies were handled individually by combining the results of participants within a study. However, effect sizes within the included studies were calculated in a fixed effect manner since the participants within a study were combined in a single value (see an overview for effect types in single-case research, [DeHart & Kaplan, 2019]). Effect sizes were interpreted in accordance with these approaches.

We examined each single-case tier within a study to calculate all three effect size scores through a data extraction process using PlotDigitizer, a reliable and valid software

program for digitizing data (Aydın & Yassıkaya, 2022). Two researchers digitized data in each tier across studies, then exported extracted data into a Microsoft Excel file for further analysis. After digitizing data, they determined the data range in each tier, reviewing tiers visually for fractional data points and accepting the nearest whole number. Limitations of the use of effect sizes with single-case research have been noted in the literature. For example, PND does not allow for calculation of standard errors or confidence intervals. Both, however, are important to report with effect sizes (e.g., Wilkinson & APA Task Force on Statistical Inference, 1999). A concern has been raised about Tau-*U* not accounting for data dependence with SCRDS. Caution also has been raised about setting and following effect size interpretations (e.g., Thompson, 2007; Vannest & Ninci, 2015).

Determination of an Evidence Base for Using the GG Procedure

We evaluated studies as *MS* and *MS-R* together against the criteria for evidence-based practices recommended by Kratochwill et al. (2013). Then, we used Horner et al.'s (2005) recommendation of three criteria for defining an evidence-based practice: (a) minimum of *five* studies categorized as *MS* and *MS-R*, (b) practice conducted by at least *three* groups of researchers with no overlapping authorship from three different geographic regions, and (c) total number of participants in combined studies equaling at least 20.

Reliability

We conducted six reliability analyses in the study: (a) inclusion and exclusion criteria, (b) quality indicators, (c) visual analysis, (d) descriptive analysis, (e) digitized data, and (f) effect size. We used a point-by-point method to determine percentage of inter-rater reliability by dividing number of agreements by total number of agreements + disagreements and multiplying by 100. Two researchers coded and analyzed 100% of the studies.

To review inclusion and exclusion criteria, as explained earlier, researchers coded studies for inclusion and exclusion, and, for disagreements, the senior researchers discussed inconsistent coding until 100% consensus was reached. For quality indicators, two researchers independently coded all studies ($n = 27$) and obtained 99.99% (range = 90%–100%) agreement. They also performed visual analyses independently for 11 studies coded as *MS* or *MS-R* and obtained 98.76% (range = 93.75%–100%) agreement. Once evaluation of studies by Kratochwill et al.'s (2013) criteria was completed, two researchers independently coded these 11 studies for descriptive analyses and obtained 98.76% (range = 95.65%–100%) agreement. If disagreement occurred between researchers, they

re-examined coded articles with a senior researcher and achieved consensus on each parameter of quality indicators and descriptive analysis. Two researchers also digitized data points in all tiers across studies for effect size analysis. They determined 1,134 data points to be digitized and disagreed on eight data points. If an error occurred using PlotDigitizer (i.e., if mouse cursor was slightly off midpoint of data point, rounding error could change value of data point), researchers operationalized agreement as the value of two data points being identical or one unit apart. Researchers obtained 99.3% (range = 96.9%–100%) agreement for using PlotDigitizer. Researchers also conducted reliability analyses for effect size for 52 tiers in the seven studies retained for effect size calculations ($n = 156$). Reliability analysis for calculation of PND resulted in 100% agreement and 99.4% (range = 88.3%–100%) agreement for Tau-*U* and IRD analyses.

Results

Quality Indicators of Single-Case Studies

As shown in Supplemental Figure 1, we located 27 studies that met the study's criterion to be included using design standards recommended by Kratochwill et al. (2013). Data on the design standards of SCRDS studies and classification of evidence of effectiveness are found in Supplemental Tables 1 and 2, respectively. Of 27 studies, we rated five studies (18.52%) as *MS*, six studies (22.22%) as *MS-R*, and 16 studies (59.26%) as *nMS*. The most common reason we did not rate studies as *MS* or *MS-R* was the lack of data points per condition. Twenty studies (74.07%) failed to obtain five data points, and seven studies (25.93%) failed to obtain three data points per condition. The second reason was not showing at least three demonstrations of effect in eight studies (29.63%). Other reasons were the failure to collect interobserver agreement for 20% of sessions in each condition ($n = 5$; 18.52%) and interobserver agreement data being lower than 80% ($n = 1$; 3.70%). Moreover, eight studies (29.63%) did not meet two indicators, four studies (14.81%) did not meet three indicators, and one study (3.70%) did not meet four indicators in the rubric. Last, although Kratochwill et al.'s (2013) rubric does not require consideration of treatment integrity data for assessing standards of studies, researchers noted that 12 studies (44.44%) failed to obtain treatment integrity data for at least 20% of each condition and, in 11 (40.74%) studies, reported treatment integrity data as being lower than 80% (see Supplemental Table 1).

Visual analysis findings among 11 studies either rated as *MS* or *MS-R* showed seven studies (63.64%) were classified as having *moderate* or *strong evidence* and four studies (36.36%) as having *no evidence* under classification of evidence of effectiveness. Of seven studies, four (57.14%) were classified as having *strong evidence*, and

the remaining three studies (42.86%) classified as having *moderate evidence* because at least one participant did not meet at least one parameter in visual analysis. We included 11 studies that met quality indicators recommended by Kratochwill et al. (2013) in the descriptive analysis and seven studies that classified as *either moderate* or *strong evidence* via visual analysis in effect size calculations (see Supplemental Table 2).

Descriptive Analysis of the GG Studies

We included 11 studies that met the design standards recommended by Kratochwill et al. (2013) in the descriptive analysis. Demographic, procedural, and outcome characteristics of the studies are shown in Supplemental Table 3. Of 11 studies, seven (63.64%) were peer-reviewed journal articles, and 4 (36.36%) were unpublished theses and dissertations.

Demographic Characteristics of Included Studies. The reviewed studies included a total of 39 participants: 33 males ($n = 7$, 63.64%) and four females ($n = 3$, 27.27%). Sex was not identified for two participants (9.09%) in one study (Brown, 2008). Twenty-three participants (58.97%) were school age (7–15 years), 12 (30.77%) were preschool age (2–6 years), and four (10.26%) were adolescents/young adults. The effects of the GG procedure were examined predominantly with individuals with ASD ($n = 20$; 51.28%) in six studies, with the second group being learners with multiple disabilities ($n = 19$; 48.72%) in five studies. Individuals with multiple disabilities were learners identified with ID and an additional disability category.

Investigators used the GG procedure to teach discrete skills in seven (63.64%) studies and chained skills in four (36.36%) studies. Discrete skills included receptive language ($n = 1$), imitation ($n = 2$), on-schedule ($n = 2$), discrimination ($n = 1$), and fine motor ($n = 1$). Chained skills included self-care ($n = 1$), daily living ($n = 1$), play ($n = 1$), and leisure ($n = 1$).

Investigators did not define mastery criteria in five studies (45.45%), but they defined criteria in the remaining six studies (54.55%). Across these studies, investigators defined criteria as 100% independent correct responses for three consecutive sessions in three studies (50%), 90% correct responding in one study (16.67%), 80% unprompted correct responding on 3 out of 5 sessions in one study (16.67%), and 75% independent correct on at least 50% of steps for four consecutive sessions in one study (16.67%).

Investigators examined the effects of the GG procedure in various settings across studies. Settings included school ($n = 5$, 45.46%), university unit ($n = 2$, 18.18%), home or university unit ($n = 1$, 9.09%), community-based home ($n = 1$, 9.09%), prevocational site ($n = 1$, 9.09%), and educational institution ($n = 1$, 9.09%). Across studies, investigators

examined the effects of the GG procedure predominantly in a one-on-one teaching arrangement ($n = 8$, 72.73%); however, they also used a group teaching arrangement ($n = 3$, 27.27%).

Procedural Characteristics of Included Studies. Of 11 studies, 7 (63.64%) used various demonstration SCRDs, 3 (27.27%) used comparison SCRDs, and 1 used a combination of demonstration and comparison SCRDs (i.e., alternating treatments design; 9.09%). Supplemental Table 2 shows specific designs used across studies.

Investigators conducted dependent variable reliability analysis in all studies ($n = 11$) and independent variable reliability analysis in the majority of studies ($n = 9$; 81.82%; e.g., Horsman, 2018). They did not address independent variable reliability in two studies (18.18%; e.g., MacDuff et al., 1993). They reported dependent and independent variable reliability as over 80% agreement.

Intervention Description. To be included in the analysis, investigators needed to implement the intervention using prompts faded using moment-by-moment decisions. Researchers analyzed GG procedures in terms of prompts used, moment-by-moment decision-making, number of trials per session, type and schedule of reinforcement, and interventionist.

Prompts Used and Moment-by-Moment Decision-Making. All studies used a physical prompt with various methods to fade the prompt. In two (18.18%) studies, investigators faded the physical prompt by changing location on the arm where the prompt was delivered (e.g., hand to wrist), and, in five (45.46%) studies, investigators faded the physical prompt with a combination of changing location on arm where the prompt was delivered (e.g., hand to elbow to upper arm to shoulder) plus fading intensity. Intensity of the prompt was faded by changing how much the investigator's hand came in contact with the learner (e.g., full hand to thumb and finger) and firmness of touch (e.g., moving to light touch). In two studies (18.18%), investigators faded the prompt by changing prompt type (e.g., physical to verbal), and, in three (27.27%) studies, investigators mentioned shadowing.

Investigators predominantly used moment-by-moment decision-making for fading prompting in 10 (90.91%) studies. In one study, investigators described prompts used and named the procedure GG; however, they did not provide information related to moment-by-moment decision-making (Paisey et al., 1989).

Number of Trials. In five (45.46%) studies, investigators identified number of trials per session but did not identify number of trials in the remaining six (54.54%) studies. When specified, they used 20 instructional trials ($n = 2$; 40%), 5 trials ($n = 2$; 40%), and 9 trials ($n = 1$; 20%) per session.

Type of Reinforcer. In nine (85.71%) studies, investigators described the type of reinforcement. They used only one type of reinforcer in three (33.33%) studies. Of these, natural reinforcers were used in (66.67%) studies, and tokens were used in one study.

Schedule of Reinforcement. Investigators identified the schedule of reinforcement in eight (72.73%) studies. Five (62.50%) studies used continuous reinforcement, one (12.50%) used continuous and fixed reinforcement, one (12.50%) used continuous and variable reinforcement, and one (12.50%) used fixed reinforcement schedules.

Interventionist. Investigators or external implementers delivered intervention in the majority of studies ($n = 7$; 63.64%). Teachers served as implementers in the remaining studies ($n = 4$; 36.36%).

Intensity and Duration of the Intervention. Six (54%) studies reported the intensity of intervention ranging from 3 to 5 days per week 3 to 5 times per day. Duration of intervention was reported in six (54%) of studies, ranging greatly from six sessions to 6 months.

Social Validity. The GG procedure was found to be socially valid in the five (45.45%) studies. They collected data from teachers in three (60%) studies, parents of participants in two (40%) studies, instructional staff in one study, therapists in one study, and a teaching assistant in one study. Investigators collected data via interviews in three (60%) studies and questionnaires in two (40%) studies.

Maintenance and Generalization. Investigators did not assess maintenance or generalization in five (45.45%) studies. In five (45.46%) studies, they assessed maintenance effects across a total of 16 participants. Duration of time that passed prior to conducting maintenance sessions varied from 1 to 36 weeks. One (20%) study did not state duration.

In seven (63.64%) studies, investigators addressed generalization of the GG procedure across a total of 22 participants, including generalization across stimuli for 10 participants ($n = 3$; 42.86%), across people for nine participants ($n = 3$; 42.86%), across materials for seven participants ($n = 2$; 28.57%), and across settings for seven participants ($n = 2$; 28.57%). Investigators did not address generalization in four (36.36%) studies.

Outcome Characteristics of Included Studies. Six (54.56%) studies reported the GG procedure was effective across 20 participants, and two (18.18%) studies revealed it was effective across all but one participant. In one study, it was effective for one out of five participants. One study reported participants did not meet criterion but acquired target

behaviors to a certain extent. Effects of the GG procedure varied in one study.

Determination of an Evidence-Based Practice

Results of this review show that the GG procedure can be considered evidence-based for teaching individuals with disabilities. First, criterion requiring a minimum of five studies categorized as *MS* and *MSR* was met in that 11 studies (i.e., Brown, 2008; Cattik & Odluyurt, 2017; de Perignat Lane, 1986; MacDuff et al., 1993; Sabielny, 2013) had acceptable rigor to support the GG procedure. Second, criterion requiring that studies be conducted by at least three research groups with no overlapping authorship from three geographical regions was met in that five of these studies were conducted by five different research groups from Turkey and different regions in the United States. Third, criterion requiring that results be demonstrated across a minimum of 20 participants (e.g., those having ASD or ID) was met in that results in the same five studies were demonstrated across 22 participants.

Effects of the GG Procedure

We determined the effects of the GG procedure by using three effect size calculations: (a) PND, (b) Tau-*U*, and (c) IRD. We applied these procedures to the seven studies that met classifications of *MS* and *MSR* and were classified either as having moderate or strong effect of evidence using the criteria recommended and defined by Kratochwill et al. (2013). Supplemental Table 4 shows PND, Tau-*U*, and IRD scores calculated across the seven studies using baseline-intervention comparisons as well as number of tiers analyzed. Supplemental Figure 2 also shows a forest plot for the studies for aggregated average Tau-*U* values with CI_{95} . PND results from baseline-intervention comparison suggest the GG procedure was *very effective* in six studies (85.71%) and *effective* in one study (14.29%). Tau-*U* results from baseline-intervention comparisons show that the GG procedure had a *strong effect* with a large CI_{95} range (except for MacDuff et al., 1993) in five studies (71.43%), and *medium to high effect* with a large CI_{95} range in two studies (28.57%). IRD results from baseline-intervention comparison suggest that seven studies had *large effect*.

Discussion

Based on the results, the GG procedure, as a stand-alone intervention, appears evidence-based practice for teaching skills to individuals with disabilities. In accordance with WWC guidelines (Kratochwill et al., 2013), analysis showed a sufficient number of studies judged to *meet design standards* or *meet design standards with reservations* ($n = 11$). The studies were conducted by at least three

independent research teams and geographic regions both within the United States and Turkey ($n = 5$), and results were demonstrated in more than 20 participants ($n = 39$) across investigations. Across the 11 studies, participants from preschool through adulthood were included, both chained and discrete skills were taught, and a variety of disability categories were represented (e.g., developmental disability, ASD, ID, multiple disability, physical disability). In addition, there were sufficient studies ($n = 7$), regions ($n = 5$) and participants ($n = 22$) to show the GG procedure to be evidence-based specifically for students with ASD and ID. The researchers conducted a meta-analysis using three effect size calculations, showing the GG procedure to be *very effective* or *effective* using PND, to have a *strong effect* or *medium to high effect* using Tau-U, and to have a *large effect* using IRD.

It appears that, based on the studies reviewed, the GG procedure had a strong effect; however, we had a limited number of studies that met standards or met standards with reservations. Given limitations of the use of effect size with SCRDS, findings should be interpreted with caution. Also, four studies showed no evidence. The review also showed, of studies reviewed, more had no effects than strong effects.

We note that, although analyzed studies produced positive effects, it is likely there are studies conducted with GG that did not produce such effects and, therefore, were not published. Shadish et al. (2016), Gage et al. (2017), and other researchers have pointed out different types of biases (e.g., selective reporting, publication bias) that may have influenced this meta-analysis and skewed the findings, given the criteria of the WWC guidelines. Gage et al. found that meta-analyses that included gray literature were less likely to reflect publication biases and that publication studies had larger effect sizes than gray literature. Therefore, because gray literature (i.e., theses, dissertations) was included in this meta-analysis, results are perhaps less biased and do not reflect inflated effect sizes to the extent they may have if only published studies had been included. Greater emphasis in special education is being placed on open science practices to make research findings more trustworthy (Cook et al., 2018), including publishing rigorous research that does not result in positive effects (Coyne et al., 2016). These changes will likely affect future establishment of evidence-based practices and guidelines for determining studies rated as *no evidence*.

In reviewing the GG studies, we collected data on two variables related to treatment integrity (i.e., 20% of all sessions with integrity percentages of at least 80%). Although these indicators are not included in the design standards by Kratochwill et al. (2013), we considered it important to report these indicators as treatment integrity data provide evidence that procedures were carried out as written and changes observed in the experiment were a result of a change in participant behavior and not a change in the

behavior of the experimenters (Barton et al., 2018). Of the 11 studies included in the meta-analysis, 8 (72.73%) included variables related to treatment integrity, thus increasing our confidence that the procedures reviewed were conducted as described in the study. Instructors should feel confident that the GG procedure is likely to be effective when used with individuals with autism and ID when conducting the procedure with fidelity.

Although results of the meta-analysis indicate the GG procedure is effective and evidence-based, we identified only 11 rigorous studies in the literature, even though no specific starting date was included in our search and despite GG being present in the literature for decades (Wolery et al., 1992). This leaves room for future researchers to continue to strengthen the evidence base through replications with other skills, age groups, disabilities, and instructional parameters to specifically determine under what conditions and for what behaviors the GG procedure is evidence-based using contemporary evidence standards. This meta-analysis showed the GG procedure has been used most commonly to teach individuals with ASD, ID, and multiple disability. Although a variety of skills have been taught with GG, further investigation is needed to identify specific skills for which GG is evidence-based as sufficient studies do not exist in a skill area to make this determination. Research is needed to identify effectiveness across ages of participants and settings as well as providing evidence that instructors can implement the procedures with fidelity.

We encountered several limitations while conducting the analysis. First, identifying GG studies that met the inclusion criteria was challenging because investigators adopted various definitions of the procedure and provided various levels of detail. Some investigators provided more specific information than others in terms of prompt used, response interval, and prompt fading, making identification and analysis of the procedures difficult. Investigators should be technologically clear when reporting methods, thereby promoting replicability. Second, we did not analyze between studies heterogeneity in this study. We recommend future researchers conduct between study heterogeneity and sensitivity analysis while studying meta-analytic study about the GG procedure. Third, investigators often used GG as a component of a treatment package (e.g., implementing GG after time delay provided for response, Cosbey & Johnston, 2006). When this occurred, we excluded the study from analysis, thereby potentially failing to capture the potential of the procedure when used in combination with other strategies and resulting in a relatively smaller number of studies. Fourth, we did not weight the score of standards in the study. In quantitative meta-analyses, effect sizes can be calculated with weighting the number of participants so that studies with more participants have a stronger “impact” on overall mean effect. We recommend future researchers conduct

effect size analyses by weighting the score of standards. Fifth, at the time of our analysis, we used the most contemporary version of the WWC standards available (Kratochwill et al., 2013). Since then, the latest WWC 5.0 version (WWC, 2022) has been released, and we do not know if similar results would have been obtained if this version had been used; however, it is likely findings would change. In one example, the 5.0 standards require at least six data points to be used in the baseline condition of a multiple baseline design to meet standards without reservations which is a change from five data points required in the previous version. It should be noted that multiple evaluation tools exist when synthesizing SCRDs (e.g., Single-Case Analysis and Review Framework [Ledford et al., 2016]; Council for Exceptional Children [Cook et al., 2015]; Version 5.0 of WWC Standards [WWC, 2022]), and the evaluation tool used may affect the outcome of findings used by future researchers. Zimmerman et al. (2018) found the evaluation tool used to synthesize research on sensory-based interventions resulted in variability of findings. Sixth, although our review of the literature ended in 2020, we reviewed enough studies since the first quarter of 2021 to make recommendations that the procedure be considered evidence-based; however, future studies will need to review research conducted with the procedure since that time. Seventh, although we included dissertations and unpublished studies to reduce the possibility of publication bias, we did not statistically assess publication bias in the study. We recommend future researchers to assess publication bias statistically. Despite these limitations, we recommend the GG procedure be considered evidence-based based on the rigor of our analysis. However, we also recommend future researchers continue to evaluate practices as new standards are developed while also recognizing studies are often designed based on standards existing at the time and should not be cast aside as irrelevant for this reason. Sixth, we did not analyze moderators by age or disability type, but we recommend future investigators address this. Last, interrater agreement was low; we noted there were different definitions for GG and, in some studies, procedures were not explicitly defined. This was an issue with descriptions by authors of studies that did not make procedures clear. We recommend future investigators improve description of procedures.

Conclusion: Implications for Research and Practice

In conclusion, although it is clear additional research is needed, researchers and practitioners should have confidence the GG procedure can be effective in teaching learners with ASD and ID. This meta-analysis adds to the literature by identifying another response prompting

procedure determined to be evidence-based just as other prompting procedures have been shown to be effective (i.e., constant time delay [Browder et al., 2009], simultaneous prompting [Tekin-Iftar et al., 2019], system of least prompts [Shepley et al., 2018]). The specific identification of evidence-based procedures, including for whom and under what conditions they are effective, provides guidance for researchers in making unbiased recommendations and for practitioners in selecting practices to use.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by Anadolu University Research Fund (Project Number: 2105E134).

ORCID iD

Elif Tekin-Iftar  <https://orcid.org/0000-0001-5512-616X>

Supplemental Material

Supplemental material for this article is available at <https://doi.org/10.1177/00224669231200461>

References

- References with one asterisk were studies reviewed for standards, two asterisks were retained for descriptive analyses, and three asterisks were retained for effect size analyses.
- *Akers, J. S., Higbee, T. S., Pollard, J. S., Pellegrino, A. J., & Gerencser, K. R. (2016). An evaluation of photographic activity schedules to increase independent playground skills in young children with autism. *Journal of Applied Behavior Analysis, 49*(4), 954–959. <https://doi.org/10.1002/jaba.327>
 - Akmanoglu, N., Yanardag, M., & Batu, E. S. (2014). Comparing video modeling and graduated guidance together and video modeling alone for teaching role-playing skills to children with autism. *Education and Training in Autism and Developmental Disabilities, 49*(1), 17–31.
 - Aydın, O., & Yassıkaya, M. Y. (2022). Validity and reliability analysis of the PlotDigitizer software program for data extraction from single-case graphs. *Perspective on Behavior Science, 45*(1), 239–257. <https://doi.org/10.1007/s40614-021-00284-0>
 - Azrin, N. H., & Foxx, R. M. (1971). A rapid method of toilet training the institutionalized retarded. *Journal of Applied Behavior Analysis, 4*(2), 89–99. <https://doi.org/10.1901/jaba.1971.4-89>
 - Barnard-Brak, L., Watkins, L., & Richman, D. M. (2021). Autocorrelation and estimates of treatment effect size for single-case experimental design data. *Behavioral Interventions, 36*(3), 595–605. <https://doi.org/10.1002/bin.1783>

- Barton, E. E., Meadan-Kaplansky, H., & Ledford, J. (2018). Independent variables, fidelity, and social validity. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 133–156). Routledge.
- *Brodhead, M. T., Higbee, T. S., Pollard, J. S., Akers, J. S., & Gerencser, K. R. (2014). The use of linked activity schedules to teach children with autism to play hide-and-seek. *Journal of Applied Behavior Analysis*, 47(3), 645–650. <https://doi.org/10.1002/jaba.145>
- Browder, D. M., Ahlgrim-Dezell, L., Spooner, F., Mims, P. J., & Baker, J. N. (2009). Using time delay to teach literacy to students with severe developmental disabilities. *Exceptional Children*, 75(3), 343–364. <https://doi.org/10.1177/001440290907500305>
- ***Brown, A. K. (2008). *Teaching children with autism which responses to imitate in an ordinary environment* (Publication No. 3325412) [Doctoral dissertation, City University of New York]. ProQuest Dissertations and Theses Global.
- *Bryan, L. C., & Gast, D. L. (2000). Teaching on-task and on-schedule behaviors to high-functioning children with autism via picture activity schedules. *Journal of Autism and Developmental Disorders*, 30(6), 553–567. <https://doi.org/10.1023/A:1005687310346>
- Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification*, 28(2), 234–246. <https://doi.org/10.1177/0145445503259264>
- ***Cattik, M., & Odluyurt, S. (2017). The effectiveness of the smart board-based small-group graduated guidance instruction on digital gaming and observational learning skills of children with autism spectrum disorder. *Turkish Online Journal of Educational Technology*, 16(4), 84–102.
- Chen, L. T., Wu, P. J., & Peng, C. Y. J. (2019). Accounting for baseline trends in intervention studies: Methods, effect sizes, and software. *Cogent Psychology*, 6(1), 1–20. <https://doi.org/10.1080/23311908.2019.1679941>
- Collins, B. C. (2022). *Systematic instruction for students with moderate and severe disabilities* (2nd ed.). Paul H. Brookes.
- Cook, B. G., Buysse, V., Klingner, J., Landrum, T., McWilliam, R., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education*, 36(4), 220–234. <https://doi.org/10.1177/0741932514557271>
- Cook, B. G., & Cook, S. C. (2011). Unraveling evidence-based practices in special education. *Journal of Special Education*, 47(2), 71–82. <https://doi.org/10.1177/0022466911420877>
- Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., & Therrian, W. J. (2018). Promoting open science to increase the trustworthiness of evidence in special education. *Exceptional Children*, 85, 104–118. <https://doi.org/10.1177/001440291879>
- Cosbey, J. E., & Johnston, S. (2006). Using a single-switch voice output communication aide to increase social access for children with severe disabilities. *Research and Practice for Persons with Severe Disabilities*, 31(2), 144–156. <https://doi.org/10.1177/154079690603100207>
- Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education. *Remedial and Special Education*, 37, 244–253. <https://doi.org/10.1177/0741932516648463>
- DeHart, W. B., & Kaplan, B. A. (2019). Applying mixed-effects modeling to single-subject designs: An introduction. *Journal of the Experimental Analysis of Behavior*, 111(2), 192–206. <https://doi.org/10.1002/jeab.507>
- ***de Perignat Lane, G. M. (1986). *Comparison of manual guidance only and manual guidance plus verbal prompts strategies with students who are blind and severely mentally retarded* (Publication No. 8908098) [Doctoral dissertation, The Johns Hopkins University]. ProQuest Dissertations and Theses Global.
- **Eliçin, Ö., & Tunalı, V. (2016). Effectiveness of tablet computer use in achievement of schedule-following skills by children with autism using graduated guidance. *Education & Science*, 41(183), 29–46. <https://doi.org/10.15390/eb.2016.5358>
- ***Gadaire, D. M., Creel, K. L., & Noto, J. (2020). Effects of group activity schedule training on preference for social play among children with autism. *Learning & Motivation*, 71, 1–11. <https://doi.org/10.1016/j.lmot.2020.101638>
- Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, 83(4), 428–445. <https://doi.org/10.1177/0014402917691016>
- *Giurlando, H. L. F. (2011). *A comparison of video modeling and graduated guidance* (Publication No. 1490996) [Master's dissertation, Northeastern University]. ProQuest Dissertations and Theses Global.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practices in special education. *Exceptional Children*, 71(2), 165–179. <https://doi.org/10.1177/001440290507100203>
- **Horsman, J. (2018). *Using graduated guidance to teach imitation of manual signs to children with intellectual disabilities* [Master's thesis, University of Kentucky]. <https://doi.org/10.13023/ETD.2018.102>
- **Ivy, S. E., Hatton, D. D., & Wehby, J. H. (2018). Using graduated guidance to teach spoon use to children with severe multiple disabilities including visual impairment. *Research and Practice for Persons with Severe Disabilities*, 43(4), 252–268. <https://doi.org/10.1177/0741932518808519>
- Jimenez, B. A., & Alamer, K. (2018). Using graduated guidance to teach iPad accessibility skills to high school students with severe intellectual disabilities. *Journal of Special Education Technology*, 33(4), 237–246. <https://doi.org/10.1177/0162643418766293>
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, 34(1), 26–38. <https://doi.org/10.1177/0741932512452794>
- ***Kurt, O. (2011). A comparison of discrete trial teaching with and without gestures/signs in teaching receptive language skills to children with autism. *Educational Sciences: Theory and Practice*, 11(3), 1436–1444.
- Ledford, J., Lane, J. D., & Barton, E. E. (2019). *Methods for teaching in early education*. Routledge.
- Ledford, J., & Severini, K. E. (2019). Using basic single case designs. In R. C. Pennington (Ed.), *Applied behavior analysis for everyone*. AAPC Publishing.

- Ledford, J. R., Lane, J. D., Zimmerman, K. N., Chazin, K. T., & Ayres, K. A. (2016). *Single case and review framework (SCARF)*. <https://ebip.vkcsites.org/scarf/>
- *Lobato, D., & Tlaker, A. (1985). Sibling intervention with a retarded child. *Education and Treatment of Children*, 8, 221–228.
- *Luiselli, J. K. (1988). Improvement of feeding skills in multi-handicapped students through paced-prompting interventions. *Journal of the Multihandicapped Person*, 1(1), 17–30. <https://psycnet.apa.org/doi/10.1007/BF01110553>
- ***MacDuff, G. S., Krantz, P. J., & McClannahan, L. E. (1993). Teaching children with autism to use photographic activity schedules: Maintenance and generalization of complex response chains. *Journal of Applied Behavior Analysis*, 26(1), 89–97. <https://doi.org/10.1901/jaba.1993.26-89>
- *McLay, L., Schäfer, M. C., van der Meer, L., Couper, L., McKenzie, E., O'Reilly, M. F., Lancioni, G. E., Marschik, P. B., Sigafoos, J., & Sutherland, D. (2017). Acquisition, preference and follow-up comparison across three AAC modalities taught to two children with autism spectrum disorder. *International Journal of Disability, Development and Education*, 64(2), 117–130. <https://doi.org/10.1080/1034912X.2016.1188892>
- McLeskey, J., Barringer, M.-D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M., Lewis, T., Maheady, L., Rodriguez, J., Scheeler, M. C., Winn, J., & Ziegler, D. (2017, January). *High-leverage practices in special education*. Council for Exceptional Children & CEEDAR Center.
- **Paisey, T. J., Whitney, R. B., & Moore, J. (1989). Person-treatment interactions across nonaversive response-deceleration procedures for self-injury: A case study of effects and side effects. *Behavioral Interventions*, 4(2), 69–88. <https://doi.org/10.1002/bin.2360040202>
- Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education*, 40(4), 194–204. <https://doi.org/10.1177/00224669070400040101>
- Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single case research: Nonoverlap of all pairs (NAP). *Behavior Therapy*, 40(4), 357–367. <https://doi.org/10.1016/j.beth.2008.10.006>
- Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, 75(2), 135–150. <https://doi.org/10.1177/001440290907500201>
- Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, 42(2), 284–299. <https://doi.org/10.1016/j.beth.2010.08.006>
- *Pellegrino, A. J. (2018). *Promoting sociodramatic play between children with autism and their typically developing peers using activity schedules* (Publication No. 10929995) [Doctoral dissertation, Utah State University]. ProQuest Dissertations and Theses Global.
- *Realon, R. E., Favell, J. E., & Dayvault, K. A. (1988). Evaluating the use of adapted leisure materials on the engagement of persons who are profoundly, multiply handicapped. *Education and Training in Mental Retardation*, 23(3), 228–237.
- *Reese, G. M., & Snell, M. E. (1991). Putting on and removing coats and jackets: The acquisition and maintenance of skills by children with severe multiple disabilities. *Education and Training in Mental Retardation*, 26(4), 398–410.
- *Rhodes, C. L. (1998). *A comparison of two instructional procedures for teaching discrete responses within chained tasks to adolescents with profound intellectual disabilities* (Publication No. 9910366). [Doctoral dissertation, Georgia State University]. ProQuest Dissertations and Theses Global.
- ***Sabiely, L. M. (2013). *A preliminary investigation of graduated guidance* (Publication No. 3673699). [Doctoral dissertation, The Ohio State University]. ProQuest Dissertations and Theses Global.
- Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single participant research: Ideas and applications. *Exceptionality*, 9(4), 227–244. https://doi.org/10.1207/S15327035EX0904_5
- Shadish, W. R., Zelinsky, N. A. M., 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(3), 656–673. <https://doi.org/10.1002/jaba.308>
- Shepley, C., Lane, J. D., & Ault, M. J. (2018). A review and critical examination of the system of least prompts. *Remedial and Special Education*, 40(5), 313–327. <https://doi.org/10.1177/0741932517751213>
- *Sigafoos, J. (1998). Assessing conditional use of graphic mode requesting in a young boy with autism. *Journal of Developmental and Physical Disabilities*, 10(2), 133–151. <https://doi.org/10.1023/A:1022813315683>
- *Sigafoos, J., Lancioni, G. E., O'Reilly, M. F., Achmadi, D., Stevens, M., Roche, L., Kagohara, D. M., van der Meer, L., Sutherland, D., Lang, R., Marschik, P. B., McLay, L., Hodis, F., & Green, V. A. (2013). Teaching two boys with autism spectrum disorders to request the continuation of toy play using an iPad®-based speech-generating device. *Research in Autism Spectrum Disorders*, 7(8), 923–930. <https://doi.org/10.1016/j.rasd.2013.04.002>
- *Sigafoos, J., Roche, L., Stevens, M., Waddington, H., Carnett, A., van der Meer, L., O'Reilly, M. F., Lancioni, G. E., Schlosser, R. W., & Marschik, P. B. (2018). Teaching two children with autism spectrum disorder to use a speech-generating device. *Research and Practice in Intellectual and Developmental Disabilities*, 5(1), 75–86. <https://doi.org/10.1080/23297018.2018.1447391>
- Solomon, B. G., Howard, T. K., & Stein, B. N. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *Journal of Behavioral Education*, 24, 438–458. <https://doi.org/10.1007/s10864-015-9221-4>
- *Strawbridge, L. A., Drnach, M., Sisson, L. A., & Van Hasselt, V. B. (1989). Behavior therapy combined with physical therapy to promote walker use by a child with multiple handicaps. *Education and Training in Mental Retardation*, 24(3), 239–247.
- Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification*, 41(4), 427–467. <https://doi.org/10.1177/0145445516676750>
- *Taylor, N. (2018). *Using photography activity schedules to facilitate independent completion of academic tasks for young children with autism* [Master's thesis, Utah State University]. <https://digitalcommons.usu.edu/etd/7317>

- Tekin-Iftar, E., Olcay-Gul, S., & Collins, B. C. (2019). Descriptive analysis and meta analysis of studies investigating of simultaneous prompting procedure. *Exceptional Children*, 85, 309–328. <https://doi.org/10.1177/0014402918795702>
- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44, 423–432. <https://doi.org/10.1002/pits.20234>
- Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. *Journal of Counseling & Development*, 93(4), 403–411. <https://doi.org/10.1002/jcad.12038>
- Vannest, K. J., Parker, R. I., Gonen, O., & Adiguzel, T. (2016). *Single Case Research: Web based calculators for SCR analysis* (Version 2.0) [Web-based application]. Texas A&M University. <http://singlecaseresearch.org>
- What Works Clearinghouse. (2022). *Procedures and standards handbook, Version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance. <https://ies.ed.gov/ncee/wwc/Handbooks>
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. <https://doi.org/10.1037/0003-066X.54.8.594>
- Wolery, M., Ault, M. J., & Doyle, P. M. (1992). *Teaching students with moderate to severe disabilities*. Longman.
- Zimmerman, K. N., Ledford, J. R., Severini, K. E., Pustejovsky, J. E., Barton, E. E., & Lloyd, B. P. (2018). Single-case synthesis tools I: Comparing tools to evaluate SCD quality and rigor. *Research in Developmental Disabilities*, 79, 19–23. <https://doi.org/10.1016/j.ridd.2018.02.003>