Quality Review of Single-Case Studies Concerning Employment Skill Interventions for Individuals with Developmental Disabilities

Margot B. Boles
Jennifer B. Ganz
Shanna Hagan-Burke
Emily V. Gregori
Leslie C. Neely
Rose A. Mason
Dalun Zhang
Victor L. Willson

Abstract

This review analyzed the quality of 39 single-case studies and 83 individual experiments focused on teaching employment skills to individuals with developmental disabilities. Experiments were assessed and included in further analyses based on the basic design standards and evidence standards (KRATOCHWILL *et al.*, 2010, 2013). Study elements were also rated according to descriptive quality indicators indicating the level of study design, procedure replication, maintenance and generalization of skills, and procedural fidelity. Video modeling, audio cueing, visual, and prompting interventions were primarily implemented in a majority of the 38 studies and 75 experiments that passed the design and evidence standards. These interventions were then assessed according to the 5-3-20 evidence-based standard (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013). According to overall analyses, quality video modeling interventions were considered as the only evidence-based intervention according to the 5-3-20 indicators (HORNER *et al.*, 2005, KRATOCHWILL *et al.*, 2010, 2013).

Key Words: quality; employment; ASD; ID; developmental disabilities; design standards.

Many individuals with developmental disabilities (DD) have difficulty transitioning from school to a career due to deficits in communication, social interaction, and task completion (GRIGAL & DESCHAMPS, 2012; HENDRICKS, 2010; HENDRICKS & WEHMAN, 2009; WEHMEYER, 1994). These are essential skills for attaining and maintaining competitive employment, but employment outcomes of individuals with DD are relatively poor in comparison to their peers without disabilities (CARTER, AUSTIN, & TRAINOR, 2012; HANLEY-MAXWELL & IZZO, 2012; NEWMAN, WAGNER, CAMETO, & KNOKEY, 2009; NEWMAN *et al.*, 2011; SANFORD *et al.*, 2011). To address

this gap in employment outcomes, a variety of curricula and specialized interventions have been incorporated into transition programs to facilitate students' employment after high school (ALWELL & COBB, 2009; Individuals with Disabilities Education Improvement Act [IDEA], 2004). Although there has been a recent surge of research focusing on interventions for adolescents and adults with DD, analyses synthesizing this body of research on the effects of interventions to teach and improve employment skills are lacking (RUSCH & DATILLO, 2012).

Synthesizing bodies of research based on quality is integral to the advancement of researcher and practitioner knowledge of reliable and effective practices (Kratochwill *et al.*, 2013). In the field of education, assessment of the quality of interventions is guided by federal legislation included in the No Child Left Behind Act (NCLB, 2001) and IDEA (2004). Both NCLB (2001) and IDEA (2004) require and promote the use of scientifically-based research practices, and seek to assess the overall quality and effectiveness of researched interventions. NCLB (2001) outlines key components of scientifically-based research practices: (a) systematic methodological elements in observation or experimental contexts, (b) systematic procedures based on statistics for analyzing data, (c) valid and reliable measures for data collection, (d) study designs that validly measure relations between the intervention and outcomes, (e) thorough descriptions of study characteristics to allow for replication or the growth of future research, and (f) acceptance of the publication through peer-review or more intensive review processes. This 6-component evaluation schema for scientifically-based research is intentionally broad and meant to include a variety of study designs and elements (i.e., group and single-case experimental design [SCED]).

It is important to apply the components of scientifically-based practices, sometimes referred to as evidence-based practices, to SCED in order to assess the quality of applied interventions (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2013). Thorough quality analyses of SCEDs can instill confidence in quality studies' intervention effectiveness within specified contexts (KRATOCHWILL *et al.*, 2013). SCED continues to contribute to special education reform and the practices geared towards individualized instruction for individuals with disabilities (GAST & LEDFORD, 2012; HORNER *et al.*, 2005). Although SCEDs focus on individual participants, a quality analysis of multiple studies and participants can create a foundation for discerning evidence-based practices (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2013). In order to appropriately apply evidence-based practice standards to SCED, an aggregate of studies implementing a specific intervention need to meet the following criteria:

(a) at least 5 studies with high-quality designs that exhibit a functional relation between the interventions and target behaviors, (b) at least 3 different research groups (no author repeats) conducted the research at 3 separate institutions; and (c) a combination of at least 20 experiments from the included studies (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013). These three requirements (also referred to as 5-3-20) define the basic foundation of considering evidence-based practices.

In order to deem a study as high quality within the evidence-based practices qualification process, individual quality indicators must be assessed (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013). These quality indicators should be operationally defined to avoid error in consistency of quality ratings across studies (COOPER, 2010). Quality indicators can be applied in stages and *What Works Clearinghouse* (WWC) provides basic design indicators for inclusion or exclusion of possible studies (KRATOCHWILL *et al.*, 2010; NINCI *et al.*, 2015). The basic design standards include: (a) purposeful manipulation of the independent variable (IV); (b) interobserver agreement (IOA) is recorded for 20% of overall data, resulting in an overall score of at least 80% agreement; (c) three different attempts to present effect at three separate points in time; and (d) each phase contains at least 3 to 5 data points. Studies are then categorized as meeting these design standards, meeting these standards with reservations, or not meeting these standards (KRATOCHWILL *et al.*, 2010, 2013). After the exclusion of studies that do not meet the basic design standards, descriptive indicators can be applied to assess the overall quality of each study without excluding additional studies. These standards include: (a) the possibility of replication based on detail given for participant characteristics, setting characteristics, interventionist characteristics, baseline and intervention procedures, and definition and measurement of dependent variables (Council for Exceptional Children [CEC], 2014; HORNER *et al.*, 2005; PALMEN, DIDDEN, & LANG, 2012; Reichow, Volkmar, & Cicchetti, 2008; ROTH, GILLIS, & DIGENNARO-REED, 2014; WOLERY, 2013); and (b) the presence and measurement of generalization, maintenance, procedural fidelity, and social validity data (BANDA, GOGOE, & MATUSZNY, 2011; CEC, 2014; REICHOW *et al.*, 2008; WALKER, RICHTER, UPHOLD, & TEST, 2010; WOLERY, 2013).

For all studies that meet the minimum basic design standards, visual analysis is a necessary step when analyzing the quality of intervention effects (KRATOCHWILL *et al.*, 2013). Visual analysis is broken up into multiple evidence quality indicators that are then applied to the studies that either meet or meet the design standards with reservations

(KRATOCHWILl *et al.*, 2013; MAGGIN, BRIESCH, & CHAFOULEAS, 2013; NINCI *et al.*, 2015). The evidence standards include the visual analysis of (a) level, (b) trend, (c) variability, (d) immediacy of effect, (e) overlap, and (f) consistency of data patterns in similar phases seen within and between baseline and intervention phases (KRATOCHWILL *et al.*, 2010, 2013). Unfortunately, there are a lack of reviews and meta-analyses incorporating these basic design and evidence standards when assessing the quality of studies focusing on SCED and employment skills for individuals with a range of DD (NINCI *et al.*, 2015; PALMEN *et al.*, 2012; ROTH *et al.*, 2014; TAYLOR *et al.*, 2012).

Multiple reviews and meta-analyses have assessed the quality of studies using at least one of these quality indicators listed above (e.g., CEC, 2014; KRATOCHWILL *et al.*, 2013), but these reviews have not combined indicators from a variety of sources to address every quality aspect of a study under investigation (BANDA *et al.*, 2011; NINCI *et al.*, 2015; PALMEN *et al.*, 2012; ROTH *et al.*, 2014; TAYLOR *et al.*, 2012; WALKER *et al.*, 2010). There is a growing need to combine all relevant SCED quality indicators for an overall quality analysis of an entire body of literature for a variety of employment skill interventions focusing on individuals with a range of DDs (NINCI *et al.*, 2015; PALMEN *et al.*, 2012; ROTH *et al.*, 2014; TAYLOR *et al.*, 2012; WALKER *et al.*, 2010).

Currently, no meta-analyses or reviews have analyzed the quality of research on multiple types of employment skill interventions for individuals with a range of DDs using multiple quality indicator sources. A comprehensive and systematic quality analysis of SCED employment skill studies can inform special education teachers, practitioners, and researchers of promising or evidence-based interventions for individuals with DD.

The purpose of this quality analysis of SCED studies implementing employment interventions for individuals with DDs is to address gaps in the current body of research and provide a response to the following question:

1. Does the body of SCED research on employment skills for adolescent and adult populations with DDs meet minimum design and evidence standards as well as adhere to descriptive design quality indicators (i.e., CEC 2014; HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2013; MAGGIN *et al.*, 2013; REICHOw *et al.*, 2008; and WOLERY, 2013)?

Methods

Article Identification

Search procedures. An electronic database search for potential studies was conducted using an electronic search engine. The databases included: *(a) Academic Search Complete, (b) Applied Technology Full Text, (c) ERIC, (d) Education Full Text, (e) Professional Development Collection*, *(f) Psychology and Behavioral Sciences Collection, (g) Social Science Full Text, (h) Vocational and Career Collection,* and *(i) Vocational Studies Complete*. All peer-reviewed and non peer-reviewed sources were retrieved from these databases. Two groups of search terms were used as Boolean phrases (includes the word *and* in between the key search terms) when searching the databases. The first group of terms included: *autis*, Asperger*, ASD, PDD*, pervasive developmental disorder, development* disab*, low-incidence dis*, intellectual* disab*, mental* retard*,* or *multiple disab**. The second group of terms included: *employ*, career*, vocation*, employ* skill*, career skill**, *vocation* skill*,* or *job skill**. The search terms identified with an asterisk broaden the database search by including the stem of the word and any possible suffix.

Inclusion and Exclusion Criteria

Title and abstract inclusion/exclusion. The title and abstract of each retrieved source were screened using the following criteria: (a) employed a SCED, (b) included at least one participant diagnosed with DD, (c) contained one or more dependent variables that measured transition skills (i.e., employment skills, independent living skills, social skills), (d) included an intervention component as the IV, (e) reflected a journal article or dissertation, and (f) published in English. Due to the focus of this meta-analysis on employment skill interventions for individuals with DD, documents were excluded if the targeted diagnoses, IVs, and dependent variables criteria were not met. Further, it is important to search peer-reviewed and other (e.g. dissertations or theses) sources to avoid publication bias (LIPSEY & WILSOn, 2001). If there was insufficient information in the title or abstract to evaluate all inclusion and exclusion criteria, the full text of that document was reviewed.

Full text inclusion/exclusion. Following the title and abstract screening, the remaining articles were evaluated using the full-text. The full text of each article was screened

using the following criteria: (a) employed a SCED (i.e., reversal/withdrawal, alternating treatments, multiple baseline, multiple probe, multi element design); (b) contained one or more dependent variables that measured employment skills; (c) included at least one participant diagnosed with DD; (d) implemented an intervention that focused on teaching and promoting independent performance of employment skills; and (e) contained a line-graph representing skill acquisition or independent task performance data (i.e., percent of task steps performed correctly and independently or number of prompts needed to complete a task). The inclusion of a line-graph representation of data was chosen as a criterion due to the need for visual analysis of the data for the quality of design and evidence reviews.

In an effort to identify all available articles pertaining to employment skill interventions for individuals with DD, an ancestral search was also conducted. This entailed searching through the references of previously included studies. Each reference was screened based on the title, following the earlier described procedures; those references determined to reflect potential studies for inclusion were pulled and the full-text was evaluated using the full set of inclusion and exclusion criteria.

Interrater reliability (IRR). Two raters independently screened 30% of the sources retrieved after the database search for the title and abstract and full text inclusion and exclusion screenings. The articles for interrater review were randomly selected from the total number of sources to avoid selection bias. In the case of a disagreement in any of the inclusion and exclusion processes, the two raters discussed the discrepancy and reached a consensus without the need of a third rater. IRR was scored as simple percent agreement (total number of agreements divided by agreements plus disagreements and then multiplied by 100) and Cohen's kappa (COHEN, 1960).

Design Quality Indicators

Basic Design Standards. Documents that passed title/abstract and full text reviews were further reviewed to whether or not each experiment present in the study met the minimum design standards (KRATOCHWILL *et al.*, 2010, 2013). Experiments are defined as each data representation of a single-case design (KRATOCHWILL *et al.*, 2010). For example, if there were two participants in a study and a multiple baseline design across skills was conducted for each participant; two different experimental data sets were independently screened according to the basic design standards. The design standards included the following

criteria: (a) purposeful manipulation of the IV; (b) IOA recorded for 20% of overall data, resulting in an overall score of at least 80% agreement (IOA components are broken up into three individually rated standards); (c) three different attempts to demonstrate an effect at three separate points in time; and (d) each phase contained at least 3 data points. Systematic manipulation of the IV is important when assessing the functional relation by applying certain conditions purposefully (KRATOCHWILL *et al.*, 2013). IRR/IOA represents the agreement between two raters or observers when collecting data, which provides a measure for reliability (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2013); therefore, IRR/IOA needs to be measured often and across time to ensure consistent reliability of the measures**.** Demonstrating or attempting to demonstrate an effect is necessary to assessing the consistent functional relation across time and should occur at least three times over three different time periods (KRATOCHWILL *et al.*, 2013). The number of data points is important when assessing the consistency of data and if behaviors are really changing from baseline to intervention. Five data points is preferred because working with individuals usually creates natural variability in the data, which can make it difficult to assess consistency if there are less than 5 data points per phase (KRATOCHWILL *et al.*, 2013).

Each experiment was analyzed according to each design standard using the rating system presented and defined in Boles (2015a, available online). Dichotomous ratings of 0 (i.e., does not meet design standards) or 2 (i.e., meets design standards) were used to assess the purposeful manipulation of the IV and attempts to present an effect. A 3 item rating system including the scores 0, 1 (i.e., meets design standards with reservations), and 2 was used for IRR measures and the number of data points because these two standards have an accepted quality measure (i.e. 3 to 4 data points per phase) and a preferred quality measure (i.e. 5 or more data points per phase).

After each study was scored according to the design standards, an overall score was assigned to each study as a whole. If a study contained at least one experiment that met or met with reservations the design standards, the entire study was scored based on this experiment. An overall score of 0 (i.e., does not meet the overall design standards) was given if one or more of the design standards listed above were scored with a zero. An overall score of 1 (i.e., meets overall design standards with reservations) was given if at least one of the design standards listed above was scored with a 1 and all the other standards were scored as 1 or 2. Finally, an overall score of 2 (i.e., meets overall design standards) was only given if all design standards were scored with a 2.

IRR for basic design standards. The basic design standards screening was completed by two raters for 50% of the articles remaining after the title/abstract and full text inclusion and exclusion process. In the case of a disagreement when evaluating the basic design standards, the two raters discussed the discrepancy and reached a consensus. IRR was scored as simple percent agreement and Cohen's kappa (COHEN, 1960).

Descriptive Design Quality Indicators

In addition to the application of basic design standards, there were key descriptive design quality indicators applied to all studies that met or met with reservations the basic design standards (CEC, 2014; KRATOCHWILL *et al.*, 2010, 2013; HORNER *et al.*, 2005; REICHOW *et al.*, 2008; WOLERY, 2013). Descriptive design indicators are those indicators that are rated according to the description and measures of all relevant elements included in each study. The description of specific study characteristics was important to analyze because sufficient detail fosters accurate replication in future research. Replication is crucial in corroborating study effects and strengthening external validity (HORNER *et al.*, 2005). In this analysis, there were five indicators that were rated according to the level of replicability based on descriptive detail: (a) participant description, (b) setting description, (c) interventionist description, (d) baseline and intervention description, and (e) dependent variable description. This rating scale along with an overall score (i.e., *Insufficient Description*, *Minimal Description*, and *Sufficient Description*) according to the level of replicability is described in Boles (2015b, available online). In addition to the five replicability indicators, four additional descriptive design indicators based on supplementary measures or assessments in each study were included: (a) maintenance, (b) generalization, (c) fidelity, and (d) social validity. The rating system for these four indicators and overall scores (i.e., *Insufficient Measure*, *Minimal Measure*, and *Sufficient Measure*) is described in Boles (2015c, available online). All of these indicators were purely descriptive with overall quality scores, and studies were not excluded based on descriptive design indicator scores.

Participant Description. In each SCED study, the participant should be described in enough detail to promote replication of the population being targeted in future research. An operational definition of a participant should include the specific diagnosis and the assessments or process that lead to the diagnosis (HORNER *et al.*, 2005). Also, participants should be identified by age, gender, and any other relevant characteristic (i.e., IQ, skill

deficits, previous training/therapy; CEC, 2014; REICHOW *et al.*, 2008). Participant descriptions were measured using a rating scale found in Boles (2015b).

Setting Description. A thorough description of the setting is an important element to consider when replicating a study (HORNER *et al.*, 2005). Setting elements such as the materials and layout of the setting, the presence of other individuals, and the location (i.e., classroom, home, work) can impact the effects of the intervention (CEC, 2014). The setting description was measured using a rating scale found in Boles (2015b).

Interventionist Description. The characteristics of the interventionist are necessary when measuring the effectiveness of intervention implementation based on the interventionist's expertise and relationship towards the individual receiving the intervention. The interventionist description should include the interventionist's occupation and relationship to the participant (i.e., teacher, peer, sibling, parent, researcher), and the interventionist's level of expertise in implementing the intervention (CEC, 2014). The interventionist description was measured using a rating scale found in Boles (2015b).

Baseline/Intervention procedure description. The baseline and intervention procedures are necessary when assessing the steps taken to prepare for and implement an intervention. A thorough description of these procedures is necessary for accurate replication and reliable measures (HORNER *et al.*, 2005). Baseline and intervention descriptions should include a thorough description of the baseline procedures (i.e., setting, materials used, assessed behaviors, session time limit) and intervention procedures (i.e., chronological steps for implementing the intervention, the behaviors required of the interventionist, setting, materials used, session time limit; HORNER *et al.*, 2005; REICHOW *et al.*, 2008). The baseline/intervention procedure description was measured using a rating scale found in Boles (2015b).

Dependent variable description. The dependent variable is important in its role in determining the success of the intervention and the overall functional relation between the intervention and the targeted behaviors. An operational definition of the dependent variables is needed to promote a valid, reliable, and objective measure of scientific observation (HORNER *et al.*, 2005). The operational definition of target behaviors, the reasons for targeting these behaviors, and a thorough description of data collection methods for the targeted behaviors are needed for future replication (CEC, 2014; HORNER, *et al.*, 2005; REICHOW *et al.*, 2008). The dependent variable description was measured using a rating scale found in Boles (2015b).

Maintenance and generalization. Maintenance and generalization enhance external validity by providing long-term data collection and/or data collection in multiple contexts (e.g., different materials, participants, interventionists, settings; HORNER *et al.*, 2005). Studies may or may not include maintenance and generalization data, but both are important to study quality and the ongoing effects of an intervention. Maintenance is measured by assessing the progress of target skills over time either with continued implementation of the intervention or as a result of the withdrawal of the intervention (HORNER *et al.*, 2005; KAZDIN, 2011). In the case of employment skill interventions, it is very important to record maintenance data due to the goal of not only acquiring but also maintaining employment skills. Generalization is considered the measure of certain effects in novel or different contexts that may include multiple participants, settings, materials, or interventionists (HORNER *et al.*, 2005). Generalization is important when assessing quality because those participants receiving an employment intervention need to know how to apply newly acquired skills to different contexts that may occur during the transition into employment (HORNER *et al.*, 2005; REICHOW *et al.*, 2008). The rating scale and description of maintenance and generalization indicators can be found in Boles (2015c).

Fidelity. Fidelity is not always included as a measure of study quality, but these are important measures when assessing the accuracy and consistency of implementation. Procedural or treatment fidelity measures the accuracy or human error when implementing the procedures included in all conditions or only in the intervention phases (LEDFORD & WOLERY, 2013). Errors in fidelity can weaken internal validity due to the intervention being implemented over time when maturation and other variables can play a role in behavior change outside of the results of the intervention (HORNER *et al.*, 2005; LEDFORD & WOLERY, 2013; WOLERY, 2013). Procedural or treatment fidelity measures should be recorded throughout the intervention or all phases using a form of data collection that measures accuracy of implementation by the interventionist for each step included in the procedures (CEC, 2014; HORNER *et al.*, 2005; and REICHOW *et al.*, 2008). The rating scale and description of the fidelity indicator can be found in Boles (2015c).

Social Validity. Social validity is defined as the overall acceptability of the procedures and outcome measures involved in an intervention program (CARTER, 2010). Social validity is crucial in maintaining an intervention program and the effects of that program (CARTER, 2010; SCHWARTZ & BAER, 1991). Social validity should measure the (a) social significance of the dependent variables (i.e., the target behaviors are beneficial to

the participant and relevant to the context), (b) the efficiency and cost effectiveness of the intervention, (c) the significance of behavior change or intervention effects were significant according to the criteria or goals set for individual studies, (d) the satisfaction of all individuals involved regarding the procedures and outcomes, and (e) the inclusion of a natural component in the intervention (i.e., the interventionist is an individual that is present in the participant's natural setting, or the intervention is implemented in the natural setting; HORNER *et al.*, 2005; and REICHOW *et al.*, 2008; WOLERY, 1978). The rating scale and description of the social validity indicator can be found in Boles (2015c).

IRR for descriptive indicators. The descriptive quality indicator analysis was completed by two raters for 100% of the articles remaining after basic design standards inclusion and exclusion process. In the case of a disagreement when evaluating the quality indicators the two raters discussed the discrepancy and reached a consensus. IRR was scored as simple percent agreement and Cohen's kappa (COHEN, 1960).

Evidence Quality Standards

Visual analysis is crucial when analyzing the overall quality of reported effects in SCED studies (KRATOCHWILL *et al.*, 2013). It is recommended that visual analysis be conducted when assessing evidence (BROSSART, VANNEST, DAVIS, & PATIENCE, 2014; KRATOCHWILL *et al.*, 2013). Visual analysis requires the review of the main components of each experiment: (a) level, (b) trend, (c) variability, (d) immediacy of effect, (e) overlap, and (f) consistency of data patterns in similar phases seen within and between baseline and intervention phases. Level is defined as the average measure of each phase. Variability takes into account the overall consistency or inconsistency of data throughout each phase. Immediacy of effect relies on the level of the last three data points in baseline compared to the level of the first three data points in the intervention phase. Finally, the consistency of data in similar phases was analyzed based on the similarity between the level, trend, and variability seen in data sets present in similar phases (i.e., comparison of data consistency in baseline phases, A_1 and A_2 of a reversal design [ABAB]; KRATOCHWILL *et al.*, 2013). These six visual analysis components were applied to four different evidence indicators (a total of 19 different items): (a) within-phase data points, (b) overall data points, (c) overall ratio of effects to non-effects, and (d) overall evidence of effect. A rating system found in Boles (2015d) for each component of visual analysis of evidence was applied to each

experiment from included studies, resulting in studies categorized as visually presenting *No Evidence*, *Moderate Evidence*, or *Strong Evidence* (KRATOCHWILL *et al.*, 2010, 2013; MAGGIN *et al.*, 2013).

In order to analyze the experiments and overall studies for quality of evidence and possible declaration of an evidence-based practice, categorization via the type of primary intervention implemented in each experiment needs to occur. Intervention codes for each experiment were employed: (a) VM (i.e., video modeling; the use of a peer, adult, participant, or point-of-view perspective in modeling target behaviors as a video or in-vivo presentation for a variety of implementations: video modeling, video prompting, video priming, video self-modeling, point of view video modeling, adult/peer video modeling, or in-vivo modeling), (b) AC (i.e., audio cueing or audio coaching delivered to the participant via an earpiece or other device while the task is performed), (c) VIS (i.e., any static pictures, written schedules, picture schedules, or scripts that prompt a participant through a task), (d) PRMTS (i.e., most-to-least or least-to-most prompting, or any other systematic prompting system using least or most intrusive prompts as the primary intervention), (e) OTH (i.e., any intervention that does not fit the categories above or combines more than one of the specified interventions). If the experiments were scored with moderate or strong evidence, they were included in the evidence-based practice analysis based on the intervention employed and the 5-3-20 evidence-based rule.

IRR for IV codes and evidence standards. The IV coding and evidence standards analysis were completed by two raters for 100% of the articles remaining after the basic design and evidence standards inclusion and exclusion process. In the case of a disagreement when evaluating the IV codes or evidence standards, the two raters discussed the discrepancy and reached a consensus. IRR was scored as simple percent agreement and Cohen's kappa (COHEN, 1960).

Results

The overall article search from designated databases resulted in 5,821 possible articles with the removal of duplicates. These articles were analyzed using the title and abstract and full text inclusion and exclusion criteria stated above. The abstract inclusion and exclusion screenings resulted in 240 articles and full text screenings resulted in 79 articles. The basic design standards (KRATOCHWILL *et al.*, 2010, 2013) were applied to these remaining

articles and resulted in 34 articles that passed all inclusion and exclusion criteria. An ancestral search of the reference section in each of the 34 articles was performed in order to find any articles that were not included in the initial search due to the search criteria or human error. Forty-three additional articles were found during the ancestral search and these were screened based on the abstract and title, full text, and basic design standards criteria. These screenings resulted in 2 additional studies for inclusion in the meta-analysis. A total of 36 articles (39 separate studies) were analyzed using the quality indicators described above. Tables 1 to 4 in provide the final analysis of each included study based on the ratings derived from the basic design standards, descriptive design quality indicators, and the evidence quality standards indicated in Boles (2015a, 2015b, 2015c, 2015d).

IRR for overall search. The IRR agreement for the abstract and title screenings was calculated as 99% with a kappa score of 0.72. IRR agreement for the full text screening was 93% with a kappa score of 0.84. Lastly, the IRR agreement for the basic design standards screening was calculated as 96% with a kappa score of 0.89.

Basic Design Standards

As a result of the initial search and the ancestral search, a total of 89 articles were analyzed using the basic design standards and rated according to the scoring system provided by Boles (2015a), based on Kratochwill and colleagues (2010, 2013). The individual experiments that did not meet the design standards or meet them with reservations recorded IRR for less than 20% of sessions, less than 3 demonstrations of possible effect, and/or less than 3 data points in at least one phase. Table 1 presents a total of 39 studies that passed the basic design standard screening by meeting all design standards or meeting the design standards with reservations. Only 6 of the original 39 studies thoroughly m*et al*l design standards. The majority of studies (n = 33) met design standards with reservations. These 33 studies only partially met standards due to reports of IOA session totals, IOA percentage agreement, and/or the number of data points. Twenty-five out of the thirty-three studies reported 20% overall percentage of sessions in which IOA was recorded, but there was no indication of the percentage of sessions for which IOA was recorded per phase or per participant/behavior. Specificity of IOA percent agreement in each phase and each participant/behavior were also missing in 24 of the 33 studies. Lastly, 21 of the 33 studies

reported only 3 to 4 data points in at least one phase instead of the preferred 5 or more data points per phase.

Table 1
Basic Design Standards

	Study Name (Date)	IV/Intervention	IOA			Effect	Data Points	OVERALL SCORE
		Purposeful Manipulation of IV	IOA Recorded	IOA Session Totals	IOA Percentage Agreement	Attempts To Present Effect	Data Points in Each Phase	Meeting Design Standards
Studies that Meet Basic Design Standards	Cihak, Alberto, Kessler, and Taber (2004)	2	2	2	2	2	2	2
	Cihak, Kessler, and Alberto (2007)	2	2	2	2	2	2	2
	Cihak, Kessler, and Alberto (2008)	2	2	2	2	2	2	2
	Dotto-Fojut, Reeve, Townsend, and Progar (2011)	2	2	2	2	2	2	2
	Kemp and Carr (1995)	2	2	2	2	2	2	2
Studies that Meet Basic Design Standards with Reservations	Allen, Wallace, Greene, Bowen, and Burke (2010a)	2	2	1	1	2	2	1
	Allen, Wallace, Renes, Bowen, and Burke (2010b)	2	2	1	2	2	2	1
	Allen, Burke, Howard, Wallace, and Bowen (2012)	2	2	1	1	2	1	1
	Bennett, Brady, Scott, Dukes, and Frain (2010)	2	2	1	1	2	1	1
	Bennett, Rangasamy, and Honsberger (2013a)	2	2	1	1	2	2	1
	Bennett, Rangasamy, and Honsberger (2013b)	2	2	1	1	2	2	1
	Bereznak, Ayres, Mechling, and Alexander (2012)	2	2	1	1	2	1	1
	Cavkaytar (2012)	2	2	1	1	2	1	1
	Chandler, Schuster, and Stevens (1993)	2	2	2	2	2	1	1
	Chang, Kang, & Huang (2013)	2	2	1	1	2	1	1
	Connis (1997)	2	2	1	2	2	2	1
	Devlin (2008)	2	2	1	2	2	2	1
	DiPipi-Hoy, Jitendra, and Kern (2009)	2	2	1	1	2	2	1
	Goh and Bambara (2013)	2	2	1	2	2	1	1
	Hume and Odom (2007)	2	2	1	2	2	1	1
	Kelly, Wildman, and Berler (1980)	2	2	1	2	2	1	1
	Lattimore, Parsons, and Reid (2006)	2	2	2	1	2	1	1
	Lattimore, Parsons, and Reid (2008) Study 1*	2	2	2	1	1	1	1
	Lattimore, Parsons, and Reid (2008) Study 2*	2	2	2	1	2	1	1
	Lattimore, Parsons, and Reid (2009)	2	2	2	1	2	1	1
	Likins, Salzberg, Stowitschek, Lignugaris/Kraft, & Curl (1989) Study 1*	2	2	1	1	2	2	1
	Likins et al. (1989) Study 2*	2	2	1	1	2	2	1
Studies that Meet Basic Design Standards with Reservation	Martin et al. (1987)	2	2	1	1	2	1	1
	Mechling and Ortega-Hurndon (2007)	2	2	1	1	2	1	1
	Mechling and Savidge (2011)	2	2	1	1	2	1	1
	Mechling and Ayres (2012)	2	2	2	1	2	1	1
	Mitchell, Schuster, Collins, and Gassaway (2000)	2	2	1	1	2	1	1

	Study Name (Date)	IV/Intervention	IOA			Effect	Data Points	OVERALL SCORE
		Purposeful Manipulation of IV	IOA Recorded	IOA Session Totals	IOA Percentage Agreement	Attempts To Present Effect	Data Points in Each Phase	Meeting Design Standards
Studies that Meet Basic Design Standards with Reservation	Morgan and Salzberg (1992) Study 1*	2	2	1	2	2	1	1
	Morgan and Salzberg (1992) Study 2*	2	2	1	2	2	1	1
	Parson, Reid, Green, and Browning (1999)	2	2	2	1	2	2	1
	Riffel et al. (2005)	2	2	1	1	2	2	1
	Van Laarhoven, Van Laarhoven-Meyers, and Zurita (2007)	2	2	2	2	2	1	1
	Wacker, Berg, Berrie, & Swatta (1985)	2	2	2	1	2	2	1
	Wacker, Berg, Choisser, and Smith (1989)	2	2	1	1	2	1	1

Note. *2 studies were included in one article

Descriptive Design Quality Indicators

The 39 included studies were then analyzed according to the overall descriptive nature of each study design element described in Boles (2015b, 2015c). Table 2 provides the descriptive design quality scores for the participant, setting, interventionist, procedure, and dependent variable descriptions. Table 3 provides the quality scores for the maintenance and generalization phases and the fidelity and social validity measures.

Participant, setting, and interventionist descriptions. A majority of the studies were thorough when providing participant descriptions. Thirty studies provided the participant inclusion criteria, age, gender, primary and secondary diagnoses (if applicable), IQ scores, and current skill levels or prior therapy. Eight studies partially met the participant description standard by giving broader or less detail (e.g., age range instead of individual ages). Only one study did not meet the standards for participant description because each participant's gender was not reported resulting in a score of 0 (see Table 2).

In contrast to participant descriptions, a majority of studies were not as thorough when reporting the setting description. Only 11 studies provided a thorough description of the setting that included the location, materials present, and presence or absence of other individuals (related or non-related to the study). In 17 studies, the setting was only partially described by including the location and the presence of other individuals or the materials present. The setting was not sufficiently described in 11 studies because only one descriptive element (i.e., location, individuals present, or materials present) was reported.

Table 2
Descriptive Quality Standards

	Study Name (Date)	Participant Description	Setting Description	Interventionist Description	Baseline and Intervention Procedure Description	Dependent Variable Description	Overall Score
Sufficient Description	Dotto-Fojut et al. (2011)	2	2	2	2	2	2
	Likins et al. (1989) Study 1*	2	2	2	2	2	2
Minimal Description	Bennett et al. (2010)	2	1	2	1	2	1
	Bennett et al. (2013a)	2	1	1	1	2	1
	Bennett et al. (2013b)	2	1	1	2	2	1
	Bereznak et al. (2012)	2	2	1	2	2	1
	Cavkaytar (2012)	2	1	1	1	2	1
	Chandler et al. (1993)	1	2	1	2	1	1
	Cihak et al. (2004)	2	1	1	2	2	1
	Cihak et al. (2007)	1	1	1	1	2	1
	Cihak et al. (2008)	1	1	1	1	1	1
	Devlin (2008)	1	1	1	1	1	1
	DiPipi-Hoy et al. (2009)	2	1	1	2	2	1
	Goh and Bambara (2013)	2	1	1	2	2	1
	Kemp and Carr (1995)	2	1	2	2	2	1
	Lattimore et al. (2006)	2	1	2	2	1	1
	Lattimore et al. (2009)	2	1	2	1	1	1
Minimal Description	Likins et al. (1989) Study 2*	2	2	2	1	1	1
	Mechling and Ortega-Hurndon (2007)	2	2	1	2	2	1
	Mechling and Savidge (2011)	2	2	1	2	2	1
	Mechling and Ayres (2012)	2	1	1	2	2	1
	Mitchell et al. (2000)	2	1	1	2	2	1
	Parson et al. (1999)	2	2	2	1	1	1
Insufficient Description	Allen et al. (2010a)	2	2	0	2	2	0
	Allen et al. (2010b)	2	0	0	2	2	0
	Allen et al. (2012)	2	2	0	2	2	0
	Chang et al. (2013)	1	1	1	1	0	0
	Connis (1997)	0	0	1	2	1	0
	Hume and Odom (2007)	2	1	0	1	1	0
	Kelly et al. (1980)	2	0	0	0	1	0
	Lattimore et al. (2008) Study 1*	2	0	2	2	2	0
	Lattimore et al. (2008) Study 2*	2	0	2	2	2	0
	Martin et al. (1987)	1	0	1	1	1	0
	Morgan and Salzberg (1992) Study 1*	2	0	1	2	2	0
	Morgan and Salzberg (1992) Study 2*	2	0	1	2	2	0
	Riffel et al. (2005)	2	0	1	1	1	0
	Van Laarhoven et al. (2007)	2	2	0	2	2	0
	Wacker et al. (1985)	1	0	0	1	1	0
	Wacker et al. (1989)	1	0	2	1	1	0

Note *2 studies were included in one article

The interventionist description indicator was similar to the setting description in that a majority of studies received low ratings. In 21 studies, the interventionist in the study was described with a title (e.g., teacher, supervisor, trainer), but the interventionist's expertise (e.g., number of years as a teacher or prior experience implementing the intervention) was never given. In contrast, 11 studies described both important aspects of the interventionist (i.e., occupation/title and expertise). Seven studies did not provide either of these main interventionist descriptions and were therefore awarded a 0 on the rating scale (see Table 2).

Procedure and dependent variable descriptions. Procedure and dependent variable descriptions were reported by a majority of studies with more detail than the setting and interventionist descriptions above. The description of the procedures in 23 studies included an explanation of all necessary elements for both the baseline and intervention procedures (i.e., setting, materials used, session time limit, steps for implementation, and behaviors of the interventionists) with enough detail for accurate replication. In contrast, 15 studies provided only enough replicable details for either the baseline or intervention phase, did not include any indication of session length, or did not include interventionist behaviors. Only one study did not give sufficient detail for either the baseline or intervention phases, resulting in a score of 0 (see Table 2).

The dependent variable description required thorough operational definitions of the target behaviors (i.e., task analysis or detailed description of the task), the reason for targeting specified behaviors, and data collection procedures. Twenty-four studies thoroughly met the dependent variable description standard. In contrast, 14 studies partially met this standard by only reporting either sufficient operational definitions of target behaviors or providing a thorough description of data collection procedures. Only one study did not operationally define the target behaviors or give sufficient detail for data collection procedures, resulting in a score of 0 (see Table 2).

Overall scores. Each of the above quality indicators was taken into account when assigning an overall rating for each of the 39 studies. For the two studies that fully m*et al*l descriptive standards, an overall score of 2 or *Sufficient Description* was given. A total of 21 studies met or partially m*et al*l descriptive quality standards and were given an overall score of 1 or *Minimal Description*. A score of 0 for any of the quality indicators above resulted in an overall score of 0 or *Insufficient Description* for 16 studies (see Table 2).

Maintenance and generalization. All 39 studies were analyzed according to the presence of maintenance or generalization data and the quality of these measures (see Table 3). More studies implemented and reported a maintenance phase than those that implemented or reported a generalization phase. Five studies not only reported maintenance measures, but the maintenance phases included data represented by 3 or more data points and recorded more than one month after the conclusion of the intervention. In contrast, 26 studies reported maintenance data, but the data was recorded a month or less after the conclusion of the intervention and/or there were less than 3 data points in this phase. A total of 8 studies did not report any maintenance measures.

Only 5 studies reported generalization measures that occurred in both baseline and intervention sessions with a total of 3 or more data points throughout all generalization measures. In contrast, 16 reported generalization measures, but generalization data was only recorded after the intervention or there were less than 3 total data points for all generalization data. A total of 18 studies did not report any generalization measures.

Fidelity and social validity. Treatment or procedural fidelity and social validity were assessed for each of the 39 studies (see Table 3). Fidelity was reported in 22 studies. Specifically, 14 studies not only reported fidelity measures, but reported fidelity for at least 20% of overall sessions with scores of at least 80% across both baseline and intervention sessions. In contrast, 9 studies reported fidelity measures for at least 20% of sessions, but fidelity was not recorded in both baseline and intervention phases. Fidelity measures were not reported in 16 studies.

Social validity was the least reported measure compared to maintenance, generalization, and fidelity measures (see Table 3). Social validity was either not reported or only included one element of the five necessary elements to a social validity measure in 21 studies (see BOLES 2015c and Table 3). Fourteen studies only reported between two and three of the five necessary elements for social validity. A majority of the studies only reported the social significance of the target behaviors and the significance of the change in behavior according to the goals/criteria set (see BOLES, 2015c). In contrast, 4 studies reported at least four of the five necessary elements for a sufficient description of social validity.

Overall scores. Each of the quality indicators above was taken into account when assigning an overall rating for each of the 39 studies. None of the studies fully m*et al*l standards for a score of 2 or *Sufficient Measure*. In contrast, 17 studies did meet or partially met a majority of the quality indicators resulting in a score of 1 or *Minimal Measure*. A total of 22 studies did not meet a majority of the standards for a score of 0 or *Insufficient Measure.*

IRR for descriptive quality standards. The overall descriptive quality analysis for participant, setting, interventionist, procedure, and dependent variable descriptions yielded 72% IRR agreement with a kappa score of 0.65. The overall descriptive quality analysis for maintenance, generalization, fidelity, and social validity measures yielded 79% IRR agreement with a kappa score of 0.71.

Table 3
Additional Phase Quality Standards

	Study Name (Date)	Maintenance	Generalization	Fidelity	Social Validity	Overall Score
Minimal Measure	Allen et al. (2010a)	1	1	0	1	1
	Allen et al. (2012)	0	1	1	1	1
	Cavkaytar (2012)	1	1	2	2	1
	Chandler et al. (1993)	2	1	2	0	1
	Cihak et al. (2004)	1	1	2	0	1
	Devlin (2008)	1	0	1	2	1
	DiPipi-Hoy et al. (2009)	0	2	2	2	1
	Dotto-Fojut et al. (2011)	1	1	1	0	1
	Hume and Odom (2007)	1	0	1	1	1
	Kemp and Carr (1995)	1	2	0	1	1
	Lattimore et al. (2009)	2	1	0	1	1
	Mechling and Ortega-Hurndon (2007)	1	2	2	0	1
	Mechling and Savidge (2011)	1	1	2	1	1
	Mitchell et al. (2000)	1	1	2	0	1
	Morgan and Salzberg (1992) Study 1*	1	2	1	0	1
	Van Laarhoven et al. (2007)	1	0	1	1	1
	Wacker et al. (1989)	1	1	1	0	1
Insufficient Measure	Allen et al. (2010b)	1	0	0	1	0
	Bennett et al. (2010)	2	0	1	0	0
	Bennett et al. (2013a)	1	0	2	0	0
	Bennett et al. (2013b)	1	1	2	0	0
	Bereznak et al. (2012)	1	0	2	0	0
	Chang et al. (2013)	1	0	0	1	0
	Cihak et al. (2007)	1	0	2	0	0
	Cihak et al. (2008)	1	0	2	0	0
	Connis (1997)	2	0	0	0	0
	Goh and Bambara (2013)	1	0	1	0	0

	Study Name (Date)	Maintenance	Generalization	Fidelity	Social Validity	Overall Score
Insufficient Measure	Kelly et al. (1980)	0	1	0	1	0
	Lattimore et al. (2006)	0	1	0	0	0
	Lattimore et al. (2008) Study 1*	1	1	0	0	0
	Lattimore et al. (2008) Study 2*	1	1	0	1	0
	Likins et al. (1989) Study 1*	1	0	0	1	0
	Likins et al. (1989) Study 2*	0	0	0	1	0
	Martin et al. (1987)	0	0	0	0	0
	Mechling and Ayres (2012)	0	0	2	1	0
	Morgan and Salzberg (1992) Study 2*	1	2	0	0	0
	Parson et al. (1999)	1	0	2	0	0
	Riffel et al. (2005)	0	0	0	2	0
	Wacker et al. (1985)	2	1	0	0	0

Note. *2 studies were included in one article

Evidence Quality Standards

A total of 83 experiments (i.e., all single-case design data representations present in each article) were analyzed using the evidence quality standards (KRATOCHWILL *et al.*, 2010; 2013) found in Boles (2015d). Four indicators that included a total of 19 different categories were applied to the baseline phase, the intervention phase, the relation between the baseline and intervention phases, and the experiment's overall effects. Table 4 provides the evidence standard scores for all experiments that scored as *Moderate* or *Strong Evidence*.

Eight experiments that scored *No Evidence* were excluded from Table 4 because these experiments did not pass the evidence standard screening. Also, unclassified interventions that did not fit the mold of the four primary interventions or included a combination of more than one of the primary interventions (9 experiments; 7 studies) were excluded from Table 4 because the 5-3-20 evidence-based rule could not be applied to these ambiguous interventions. A total of 66 experiments were analyzed according to the 5-3-20 evidence-based rule (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013).

Table 4

Evidence Quality Indicators

Intervention Type	Study Name (Date)	Indicator #1				Indicator #2				Indicator #3							Indicator #4			
		Baseline Analysis				Within Phase Analysis				Between Phase Basic Effects							Overall Effectiveness (All Phases)			
		BA-CH	BA-PR	BA-CV	BA-TR	WP-DP	WP-PR	WP-CV	WP-TR	BW-BE	BW-IL	BW-IT	BW-CL	BW-CV	BW-OV	BW-SP	OV-DP	OV-TE	OV-ER	OV-EE
Video Modeling	Van Laarhoven et al. (2007)	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	2	2	1
		1	1	1	0	1	1	0	1	1	1	1	1	1	1		1	2	2	1
		0	1	1	0	1	1	1	1	1	1	1	1	1	1		1	2	2	1
		1	1	0	0	1	1	0	1	1	1	1	1	1	1		1	2	2	1
AC	Bennett et al. (2013a)	1	1	1	0	2	1	1	1	1	1	0	1	1	1		2	2	2	2
	Bennett et al. (2013b)	1	1	1	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
		1	1	1	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
	Allen et al. (2012)	1	1	1	0	1	1	0	1	1	1	1	1	1	1		1	2	2	1
	Bennett et al. (2010)	0	1	1	0	1	1	1	1	0	0	0	1	0	0		1	2	1	1
VIS	Cihak et al. (2004)	1	0	1	0	2	1	1	1	1	1	1	1	1	1		2	2	2	2
		1	1	0	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
	Cihak et al. (2007)	1	1	1	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
		1	1	1	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
		1	1	1	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
		1	1	1	0	2	1	1	1	1	1	1	1	1	1		2	2	2	2
	Cihak et al. (2008)	1	1	1	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
	Connis (1997)	0	0	1	0	2	1	1	1	1	1	0	1	1	1		2	2	2	2
	Devlin (2008)	1	1	1	0	2	1	1	1	1	1	1	1	1	1		2	2	2	2
	Dotto-Fojut et al. (2011)	1	1	1	1	2	1	0	0	1	1	1	1	0	1		2	2	2	2
		1	1	1	1	2	1	0	1	1	0	0	1	1	1		2	2	2	2
	Martin et al. (1987)	1	0	0	0	2	1	1	1	1	1	1	1	1	1	1	2	2	2	2
		1	0	0	0	1	1	1	1	1	0	0	1	0	0		1	2	1	1
		1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1
		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1
	Riffel et al. (2005)	0	0	0	0	2	0	0	0	1	0	0	1	0	0		2	1	1	1
PRMT	Parson et al. (1999)	1	1	1	0	2	1	0	1	1	1	1	1	1	1		2	2	2	2

Table 4
Evidence Quality Indicators

Intervention Type	Study Name (Date)	Indicator #1				Indicator #2				Indicator #3							Indicator #4			
		Baseline Analysis				Within Phase Analysis				Between Phase Basic Effects							Overall Effectiveness (All Phases)			
		BA-CH	BA-PR	BA-CV	BA-TR	WP-DP	WP-PR	WP-CV	WP-TR	BW-BE	BW-IL	BW-IT	BW-CL	BW-CV	BW-OV	BW-SP	OV-DP	OV-TE	OV-ER	OV-EE
Video Modeling	Allen et al. (2010a)	1	1	1	1	2	0	0	0	1	1	1	1	1	1		2	2	2	2
	Allen et al. (2010b)	1	1	1	1	2	0	0	0	1	0	0	1	1	0		2	2	2	2
	Morgan and Salzberg (1992) Study 1*	1	1	1	1	2	0	0	1	1	1	1	1	1	1		2	2	2	2
		1	0	0	0	2	1	0	0	1	1	0	1	1	1		2	2	2	2
	Likins et al. (1989) Study 1*	1	1	1	1	2	1	1	1	1	1	1	1	1	1		2	2	2	2
	Likins et al. (1989) Study 2*	1	1	1	0	2	1	1	1	1	0	1	1	1	1		2	2	2	2
	Bereznak et al. (2012)	1	1	1	1	1	0	0	1	1	1	0	1	1	1		1	2	2	1
	Chandler et al. (1993)	1	1	1	0	1	1	1	1	1	0	1	1	1	1		1	2	2	1
		1	1	1	0	1	1	1	1	1	0	1	1	1	1		1	2	2	1
		1	1	1	0	1	1	0	1	1	0	1	1	1	1		1	2	2	1
		1	1	1	0	1	1	1	1	1	0	0	0	1	1		1	2	2	1
	Kelly et al. (1980)	1	1	1	0	1	1	0	0	1	1	0	1	1	1		1	2	2	1
		1	1	1	1	1	0	0	0	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	1	0	0	0	1	0	0	1	1	1		1	2	2	1
	Mechling and Ortega-Hurndon (2007)	1	1	1	0	1	1	1	1	1	0	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	0	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	0	1	1	1	1		1	2	2	1
	Mechling and Saavidge (2011)	1	0	0	0	1	1	1	1	1	1	0	1	1	1		1	2	2	1
		0	0	0	1	1	1	1	1	1	1	0	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	0	1	0	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	0	1	0	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	2	2	1
	Mechling and Ayres (2012)	1	1	1	1	1	1	1	0	1	1	0	1	1	1		1	2	2	1
		1	1	1	1	1	1	0	1	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	0	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	2	2	1
	Morgan and Salzberg (1992) Study 1*	1	1	1	0	1	1	1	0	1	1	0	1	0	0		1	2	2	1
	Morgan and Salzberg (1992) Study 2*	1	0	0	0	1	1	1	1	1	1	1	1	1	1		1	2	2	1
PRMT	Lattimore et al. (2006)	1	1	0	0	1	1	1	1	1	1	1	1	1	1		1	2	2	1
		1	0	1	0	1	1	1	1	1	0	0	1	1	0		1	2	2	1
	Lattimore et al. (2008) Study 1*	1	1	1	0	1	1	1	1	1	1	1	1	1	1		1	2	2	1
	Lattimore et al. (2008) Study 2*	1	1	0	0	1	1	1	0	1	1	1	1	1	1		1	2	2	1
		1	1	1	1	2	1	1	1	1	0	0	1	1	1		2	2	2	2
	Lattimore et al. (2009)	1	1	1	0	1	1	1	1	1	1	1	1	1	1		1	2	2	1
	Hume and Odom (2007)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	2	2	1

Video modeling. The evidence standard screening for video modeling interventions included 43 experiments (15 studies). All participants in the video modeling studies were 12 years old or older and had a diagnosis of DD. Video modeling studies included 20 participants with ASD, 10 participants with ASD and ID, and 14 participants with ID. A majority of participants (n = 23) ranged from ages 16 to 21 years old. For the evidence standards, in *Indicator #1*, a majority of the video modeling baseline phases were given the highest scores

for the categories: (a) baseline data indicating a participant's need for an intervention for the targeted skill (*n* = 40 experiments), (b) baseline data indicating predictability (*n* = 37 experiments), (c) baseline data indicating stability (*n* = 35 experiments), and (d) baseline data indicating trend toward the hypothesized effect (*n* = 27 experiments). In contrast, the baseline trend standard included the largest number of experiments (*n* = 15 experiments) that received a score of 0 (i.e., the baseline trend moves in the opposite direction of the hypothesized direction).

In *Indicator #2*, the experiments' intervention phases were scored based on the number of data points in each phase, data predictability, data variability, and data trend. Scores of 0 or 1 (in the case of a 3 item rating system for the number of data points) were found most commonly in the categories of the number of data points present in each phase (i.e., 3 to 4 data points per phase; *n* = 37 experiments) and of the consistency of data variability (i.e., intervention phase data fluctuated too erratically to indicate consistency; *n* = 15 experiments). Only 6 video modeling experiments contained 5 or more data points per phase were given scores of 2.

In *Indicator #3*, the relation between the baseline and intervention phases was scored based on the basic effects between phases, immediate change in level, immediate change in trend, the overall change in level, the overall change in variability, the overall overlap of data between phases, and similarity in data phases (only applicable fore reversal designs). Scores of 0 were most commonly applied in the evidence standard categories for immediacy of change in level (i.e., no immediate change in level within the first 3 data points of the intervention phase compared to the last 3 data points of the baseline phase; *n* = 17 experiments) and the immediacy of change in trend (i.e., no change in trend within the first 3 intervention data points because the baseline phase already presented a trend toward the intervention's hypothesized direction or the variability of the data made it difficult to visually establish a trend; *n* = 19 experiments).

In *Indictor* #4, the overall evidence of effect was scored as a result of the overall number of data points, overall number of treatment effects, and overall treatment effect ratio. A majority of video modeling experiments received a score of *Moderate Evidence* (n = 30 experiments), and only 6 experiments received a score of *Strong Evidence.* In contrast, 7 experiments received a score of *No Evidence* (i.e., less than 3 treatment effects and/or less than a 3:1 ratio of treatment effects to non-effects) and excluded from the evidence-based analysis.

Evidence-based analysis. As a result of the evidence analysis, 36 video modeling experiments (13 studies) passed the evidence standards and were analyzed according to the 5-3-20 evidence-based rule (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013). Overall, video modeling interventions were implemented by 6 different research groups from separate institutions, and included more than 20 experiments, which indicates that video modeling can be considered an evidence-based intervention in teaching employment skills to individuals with DD.

Audio cueing/coaching. The evidence standard screening for audio cueing/coaching interventions included 5 experiments (4 studies). Audio cueing/coaching intervention studies included 7 participants with ASD, 2 participants with ASD with ID, and 3 participants with ID; and a majority of participants (n = 8) ranged in age from 16 to 21 years old. For the evidence standards, in *Indicator #1*, a total of 3 audio cueing/coaching experiments (3 studies), scored a 0 due to baseline data trend. For *Indicator #2*, scores of 0 or 1 were found most commonly in the categories of the number of data points present in each phase and of the consistency of data variability. A total of 2 audio cueing/coaching experiments (2 studies) received a score of 1 and 3 audio cueing/coaching experiments (2 studies) received a score of 2 for number of intervention phase data points. In addition, 1 audio cueing/coaching experiment (1 study) was scored as 0 due to variability of the intervention phase data. For *Indicator #3*, a score of 0 was most commonly applied in the evidence standard categories of immediacy of change in level (n = 1 experiment) and the immediacy of change in trend (n = 2 experiments). For *Indictor #*4, the overall evidence of effect for audio cueing/coaching interventions was scored as 1 (*Moderate Evidence*) for 2 experiments (2 studies) and 2 (*Strong Evidence*) for 3 experiments (2 studies).

Evidence-based analysis. As a result of these overall scores, a total of 5 quality audio cueing/coaching experiments (4 studies) passed the evidence standard screening. The audio cueing/coaching interventions were not determined as evidence-based interventions due to the small number of studies/experiments and overlapping author groups (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013).

Visuals. The evidence standard screening for visual interventions included 17 experiments (8 studies). Visual intervention studies included 4 participants with ASD, 0 participants with ASD with ID, and 25 participants with ID. A majority of these participants (n = 23) ranged in age from 16 to 21 years old. For the evidence standards, in *Indicator #1*, a total of 7 visual experiments (6 studies) scored a 0 due to baseline trend. For *Indicator #2*,

scores of 0 were applied most commonly in the category of consistency of data variability (*n* = 3 experiments). Scores of 2 were given to 14 experiments (10 studies) with 5 or more data points per phase. For *Indicator #3*, scores of 0 were most commonly found in the evidence standard categories of immediacy of change in level (*n* = 4 experiments) and the immediacy of change in trend (*n* = 5 experiments). In addition, 3 experiments received a score of 1 for the similarity of data patterns in similar phases (i.e., reversal designs). For *Indictor #4*, the overall evidence of effect for visual interventions was scored as 0 (*No Evidence)* for 1 experiment, 1 (*Moderate Evidence*) for 4 experiments (2 studies), and 2 (*Strong Evidence*) for 12 experiments (7 studies).

Evidence-based analysis. As a result of these overall scores, a total of 16 quality visual experiments (8 studies) passed the evidence standard screening. Visual interventions did not me*et al*l of the 5-3-20 evidence-based standards (HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013). Overall, there were 8 visual intervention studies implemented by 5 different research groups from separate institutions, but there were less than the required 20 experiments (KRATOCHWILL *et al.*, 2010, 2013), which indicates that visuals cannot be considered an evidence-based practice for employment skills for individuals with DD.

Prompting. The evidence standard screening for prompting interventions included 8 experiments (6 studies). Prompting interventions included 4 participants with ASD, 7 participants with ASD with ID, and 2 participants with ID. A majority of these participants (*n* = 12) were 22 years old or older. For the evidence standards, in *Indicator #1*, a majority of the prompting baseline phases received a score of 1 in the categories of the need for behavior change (*n* = 8 experiments), data predictability (*n* = 7 experiments), and data consistency (*n* = 6 experiments). In contrast, a total of 6 prompting experiments (4 studies) received a score of 0 due to baseline trend. For *Indicator #2*, scores of 1 were found most commonly for number of data points per phase (*n* = 6 experiments). In contrast, 2 prompting experiments (2 studies) contained 5 or more data points per phase and received a score of 2. For *Indicator #3*, a score of 0 was only given to one experiment across the evidence standard categories of immediacy of change in level and the immediacy of change in trend. Only one experiment was analyzed according to the category ratings of similar data phases and received a score of 0 due to noticeable difference in data of similar phases. In *Indictor* #4, the overall evidence of effect for prompting interventions was scored as 1 (*Moderate Evidence*) for 2 experiments (2 studies) and a 2 (*Strong Evidence*) for 6 experiments (5 studies).

Evidence-based analysis. As a result of these overall scores, a total of 8 quality, prompting experiments (6 studies) passed the evidence standard screening. The prompting interventions were not determined to be evidence-based interventions due to the overlapping author groups resulting in only two different author research groups/institutions out of the 6 total studies.

IRR for IV codes and evidence standards. The overall IRR agreement score for IV coding was 98% with a kappa score of 0.97. The overall IRR agreement score for evidence standard ratings was 91% with a kappa score of 0.81. The majority of disagreements occurred when scoring baseline data variability, baseline data trend, intervention data variability, intervention data trend, the immediacy of change in trend between the baseline and intervention phases, and the overall change in variability between phases.

Discussion

This review analyzed the quality of 79 studies implementing interventions for individuals with DD to promote independence and acquisition of a range of employment skills via the basic design standard screening. A total of 39 studies from the original 79 studies passed the basic design standard screening and were analyzed via the descriptive quality indicators and evidence standards (CEC, 2014; HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2010, 2013; MAGGIN *et al.*, 2013; REICHOW *et al.*, 2008; and WOLERY, 2013). Following the exclusion of studies via the evidence standards (KRATOCHWILL *et al.*, 2010, 2013), 38 studies were then categorized according to primary intervention implementation.

The basic design standard analysis resulted in the exclusion of 40 complete studies. The majority of these studies were excluded due to failure to meet 20% IOA across sessions, at least 3 attempts to present effect, and/or at least 3 data points per phase. The year of publication for these studies may play a role in the failure to meet the basic design standards. Thirty of these forty excluded studies were published earlier than 2005. This may be significant because in 2005, Horner and colleagues published research design standards in SCED by calling for better IOA measures, experimental control via 3 attempts to demonstrate effects at 3 different points in time, and 3 or more data points per phase. Before 2005, the standards for quality single-case research were not established, making it difficult for a

majority of studies published before 2005 to meet quality design standards. For those studies that were included in the overall quality analysis, 4 out of the 6 studies that m*et al*l design standards were published on or after 2005. Twenty-one studies out of the thirty-three studies that met the standards with reservations were published on or after 2005.

A majority of included studies (n = 33) met the design standards with reservations due to the lack of specificity when describing IOA procedures and/or only reported 3 to 4 data points per phase. Many studies do not report the percentages of recorded IOA sessions for each participant/behavior or per phase. Overall session percentages are given, which may lead the reader astray when considering just how many sessions per participant/behavior and phase that IOA observations actually resulted. IOA is important when considering the reliability of the data; therefore, it should affect how readers draw conclusions from the overall results of the study (HORNER *et al.*, 2005). The preferred minimum of 5 data points per phase is also important as it more clearly represents the predictability, consistency, and trend of the data set (HORNER *et al.*, KRATOCHWILL *et al.*, 2010, 2013).

Standards based on the descriptive nature of each of the main elements in each study were analyzed based on a combination of indicators gleaned from CEC (2014), Horner *et al.* (2005), Reichow *et al.* (2008), and Wolery (2013). A very small number of studies met the standards for all indicators addressing the description of study elements. The majority of the studies gave insufficient descriptions for the setting or interventionist. Each element of the study is important to describe thoroughly in order to promote consistent reliability and replication for future researchers who wish to implement a similar intervention and expand the literature (HORNER *et al.*, 2005). The only way to build the evidence base for SCED is to promote replicability across authors, institutions, and participants; therefore, all relevant details regarding participant characteristics, settings, interventionists, baseline and intervention procedures, and dependent variables must be at the highest standard to promote replicability (HORNER *et al.*, 2005). Further, descriptive characteristics, procedures, and outcomes can inform practitioners of effective interventions that are suited best for certain populations and precise step-by-step procedures for implementing these interventions with fidelity (REICHOW *et al.*, 2008).

Beyond baseline and intervention data, detailed and valid measures of maintenance, generalization, procedural fidelity, and social validity are needed to promote replicability and efficacy of specified interventions beyond the confines of the experimental context (CEC, 2014; HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2013; REICHOW *et al.*, 2008;

WOLERY, 2013). The quality analysis for maintenance, generalization, fidelity, and social validity verified that many studies failed to report these measures or only partially met the descriptive standards for these additional measures. None of the studies m*et al*l of the quality standards for each measure (i.e., *Sufficient Measure*); therefore, they received overall scores of *Minimal Measure* or *Insufficient Measure*. Generalization, fidelity, and social validity measures were frequently left out or insufficiently described when describing study procedures or results. This is common due to the lack of time or resources to implement generalization sessions, implement fidelity measures by an observer other than the interventionist, disperse social validity questionnaires to all stakeholders, and/or create a valid social validity measure. Generalization measures are necessary to assess the performance of target skills in a variety of contexts. To validly measure generalization, data recording needs to occur in every phase of the study to strengthen external validity and confidently measure the effects of the intervention in a different context (HORNER *et al.*, 2005). Fidelity is needed to assess the consistency of intervention implementation and how this might affect the overall results (HORNER *et al.*, 2005; WOLERY, 2013). If the intervention is not implemented with fidelity, this can create a weaker foundation for the functional relation between the intervention and the target behaviors (WOLERY, 2013). Social validity is needed to assess the social reasons behind implementing a specific intervention as well as the stakeholders' opinions about treatment acceptable, efficiency, effectiveness, and continuation of the intervention outcomes when the study is completed (HORNER *et al.*, 2005; REICHOW *et al.*, 2008; WOLF, 1978). Without social validity measures, the entire study comes into question regarding the overall beneficial nature of this intervention and target behaviors for the participant and all other stakeholders (HORNER *et al.*, 2005; REICHOW *et al.*, 2008).

Each study's experiment(s) were analyzed using the evidence standards provided by the *WWC* (KRATOCHWILL *et al.*, 2010, 2013). This is necessary when considering the visual components of the data and how this plays into the overall effects. Visual analysis is commonly used to analyze the effects of single-case research and is recommended in addition to using effect sizes (KRATOCHWILL *et al.*, 2013). All studies except one either partially met the standards (*Moderate Evidence)* or m*et al*l the standards (*Strong Evidence)*. This is encouraging when analyzing the overall effectiveness of interventions to promote employment skills because a majority of the studies that met (with or without reservations) the basic design standards also met or partially met the evidence standards. Even though some experiments received low scores for immediacy of change in level or baseline trend, the

quality of single-case designs consistently revealed obvious positive treatment effects across a majority of experiments and studies. Overall positive effects could be the result of publication bias, or only publishing studies that show significant or visually unambiguous effects, but the overall search for articles in this review included both peer-reviewed and non-peer-reviewed articles to combat this bias (COOPER, 2010). Regardless, the bias may still exist, but with the present information, 38 studies and 75 experiments were found to have moderate or strong evidence for overall positive treatment effects.

These 38 studies were categorized into groups according to the type of intervention implemented. The interventions most commonly implemented were video modeling and visuals. Only video modeling interventions met the 5-3-20 evidence-based practice standards (HORNER *et al.* 2005; KRATOCHWILL *et al.*, 2010, 2013). However, there are some limitations regarding the definitions of this intervention. Video modeling is a broad definition of this type of intervention because there are many types of models (e.g., peers, adults, self, or point-of-view, in-vivo) as well as different implementations (e.g., video priming vs. video prompting) that make it difficult to specifically determine the most effective component or variety of video modeling (BELLINI & AKULLIAN, 2007; MASON *et al.*, 2012, 2013). There were not enough studies included in our meta-analysis for a component analysis for video modeling to specify which type of component brought the strongest effects (BOLES *et al.*, 2015). Therefore, the results of this evidence-based intervention should be analyzed with caution and future research should focus on studies analyzing the effects of specific components in these interventions.

When analyzing all studies that employed one of the four interventions, there was much author or institution overlap, which made it difficult to reliably analyze similar intervention effects in different contexts. Each of the interventions that did not meet evidence-base standards (i.e., audio cueing, visuals, and prompting) should be replicated across multiple authors and institutions to ensure intervention effectiveness and social validity across contexts as well as enhance the evidence base.

This review implemented a thorough quality analysis for all of the included studies, but there were limitations. First, all data collection was scored using rating scales, which made it difficult in some cases to have a high percent of IRR agreement. Although most of the rating scales only had 2 to 3 items, attaining reliability across raters was difficult with more abstract measures such as the descriptive design indicators. Difficulties arose between the raters because there were either too many components included in each item for the specified

study element or each study reported the elements in different ways (e.g., used different jargon, reported it with different measures, reported in a different section of the paper where it was harder to find) which made it hard to discern the correct score for that study element.

Second, the lack of a sufficient number of studies that implemented audio cueing and prompting interventions made it difficult to make assumptions based on the quality and evidence of the intervention type. In addition, 7 out of the 38 studies did not fall into any intervention categories because the interventions were implemented as packages and included a wide range of intervention components. These intervention packages make it difficult to analyze the quality and efficacy of specific intervention components. Additionally, many of these studies had overlapping authors, which excludes the studies from being counted separately as additions to the evidence base for employment skill interventions.

Third, a broad spectrum of participants were included in these studies. The most common type of DD included ASD, ID, comorbidity of DD with other disabilities (e.g., ASD with ID), and multiple disabilities (i.e., included more than 2 diagnoses). Quality, effective, and evidence-based video modeling interventions included participants with the most diverse disabilities (i.e., ASD, ASD with ID, ID) and ages (i.e., 12 to 15 years old, 16 to 21 years old, and 22 years old and older). The diversity of participants in video modeling interventions made it difficult to determine this as an evidence-based practice for a specific population. Continuous research is needed regarding larger samples of participants with ASD, ASD with ID, and ID ages 12 year old and older to specify video modeling efficacy for certain individuals. Research is also needed for visual interventions regarding larger samples of participants with ASD and ASD with ID ages 12 to 15 years old and 22 years old and older.

This quality review added to the employment skill intervention literature base, but future research needs to focus on filling the existing gaps in this body of research. Researchers need to concentrate on replicating existing and promising interventions across institutions and authors for this adolescent and adult population with DD to increase the evidence base. More SCED research is needed for audio cueing/coaching and specified prompting interventions for individuals diagnosed with ASD, ASD with ID, and ID for future assessment of evidence-based practices regarding employment skills. Further, researchers need to thoroughly describe all participants, settings, implementers, procedures, and target behaviors; and include maintenance, generalization, fidelity and social validity measures according to the quality standards (CEC 2014; HORNER *et al.*, 2005; KRATOCHWILL *et al.*, 2013; MAGGIN *et al.*, 2013; REICHOW *et al.*, 2008; and WOLERY, 2013).

References

*ALLEN, K.D., WALLACE, D.P., GREENE, D.J., BOWEN, S.L., & BURKE, R.V. Community-based vocational instruction using videotaped modeling for young adults with autism spectrum disorders performing in air-inflated mascots. *Focus on Autism and Other Developmental Disabilities, 25*(3), 186-192, 2010a.

*ALLEN, K.D., WALLACE, D.P, RENES, D., BOWEN, S.L., & BURKE, R.V. Use of video modeling to teach vocational skills to adolescents and young adults with autism spectrum disorders. *Education and Treatment of Children,* 33(3), 339-349, 2010b.

*ALLEN, K.D., BURKE, R.V., HOWARD, M.R., WALLACE, D.P., & BOWEN, S.L. Use of audio cuing to expand employment opportunities for adolescents with autism spectrum disorders and intellectual disabilities. *Journal of Autism and Developmental Disorders*, 42, 2410-2419, 2012.

ALWELL, M., & COBB, B. Functional life skills curricular interventions for youth with disabilities: systematic review. *Career Development for Exceptional Individuals*, 32, 82-93, 2009.

BANDA, D.R., DOGOE, M.S., MATUSZNY, R.M. Review of video prompting studies with persons with developmental disabilities. *Education and Training in Autism and Developmental Disabilities*, 46(4), 514-527, 2011.

BELLINI, S., & AKULLIAN, J. A meta-anlaysis of video self-modeling interventions for children and adolescents with autism spectrum disorders. *Exceptional Children*, 73, 264-287, 2007.

*BENNETT, K., BRADY, M.P, SCOTT, J., DUKES, C., & FRAIN, M. The effects of covert audio coaching on the job performance of supported employees. *Focus on Autism and Other Developmental Disabilities*, 25(3), 173-185, 2010.

*BENNETT, K.D., RANGASAMY, R., & HONSBERGER, T. The effects of covert audio coaching on teaching clerical skills to adolescents with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 43, 585-593, 2013a.

*BENNETT, K.D., RANGASAMY, R., & HONSBERGER, T. Further examination of covert audio coaching on improving employment skills among secondary students with autism. *Journal of Behavioral Education*, 22, 103-119, 2013b.

*BEREZNAK, S., AYRES, K.M., MECHLING, L.C., & ALEXANDER, J.L. Video self-prompting and mobile technology to increase daily living and vocational independence for students with autism spectrum disorders. *Journal of Developmental and Physical Disabilities*, 24, 269-285, 2012.

BOLES, M. *Basic Quality Design Standards Rating Scale for Single-case Design*. Disponível em: <*http://hdl.handle.net/1969.1/153769*> 2015a.

BOLES, M. *Descriptive Quality Indicators Rating Scale for Single-case Design*. Disponível em: <http://hdl.handle.net/1969.1/153770> 2015b.

BOLES, M. *Additional Phase and Assessment Descriptive Quality Indicators for Single-case Design*. Disponível em: <http://hdl.handle.net/1969.1/153771> 2015c.

BOLES, M. *Visual Analysis/Evidence of Effect Quality Indicators for Single-case Design*. Disponível em: <http://hdl.handle.net/1969.1/153772> 2015d.

BOLES, M., GANZ, J., HAGAN-BURKE, S., HONG, E., NEELY, L., DAVIS, J., … WILLSON, V. *Effective Interventions In Teaching Employment Skills to Individuals with Developmental Disabilities: A Single-Case Meta-Analysis*. Submitted manuscript, 2015.

BROSSART, D.F., VANNEST, K.J., DAVIS, J.L., & PATIENCE, M.A. Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation*, 24, 464-491, 2014.

CARTER, S.L. *The social validity manual:* A guide to subjective evaluation of behavior interventions. San Diego, CA: Elsevier, 2014.

CARTER, E.W., AUSTIN, D., & TRAINOR, A.A. Predictors of post-school employment outcomes for young adults with severe disabilities. *Journal of Disability Policy Studies*, 23, 50-63, 2012.

*CAVKAYTAR, A. Teaching café waiter skills to adults with intellectual disability: A real setting study. *Education and Training in Autism and Developmental Disabilities*, 47, 426-437, 2012.

*CHANDLER, W., SCHUSTER, J.W., & STEVENS, K.B. Teaching employment skills to adolescents with mild and moderate disabilities using constant time delay procedure. *Education and Training in Mental Retardation*, 28, 155-168, 1993.

*CHANG, Y-J., KANG, Y-S., & HUANG, P-C. An augmented reality (AR)-based vocational task prompting system for people with cognitive impairments. *Research in Developmental Disabilities*, 34, 3049-3056.

*CIHAK, D.F., ALBERTO, P.A., KESSLER, K.B., & TABER, T.A. An investigation of instructional scheduling arrangements for community-based instruction. *Research in Developmental Disabilities*, 25, 67-88, 2004.

*CIHAK, D.F., KESSLER, K.B., & ALBERTO, P.A. Generalized use of a handheld prompting system. *Research in Developmental Disabilities*, 28, 397-408, 2007.

*CIHAK, D.F., KESSLER, K., & ALBERTO, P.A. Use of a handheld prompting system to transition independently through vocational tasks for students with moderate and severe intellectual disabilities. *Education and Training in Developmental Disabilities*, 43(1), 102-110, 2008.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46, 1960.

*CONNIS, R.T. The effects of sequential pictorial cues, self-recording, and praise on the job task sequencing of retarded adults. *Journal of Applied Behavioral Analysis*, 12, 355-361 1979.

COOPER, H. *Research synthesis and meta-analysis:* A step-by-step approach. 4 Ed. Thousand Oaks, CA: SAGE Publications, 2010.

COUNCIL FOR EXCEPTIONAL CHILDREN (CEC). *Council for Exceptional Children standards for evidence-based practices in special education.* Disponível em: <http://www.cec.sped.org/Standards/Evidence-Based-Practice-Resources-Original> 2014.

*DEVLIN, P. Enhancing the job performance of employees with disabilities using te self-determined career development model. *Education and Training in Developmental Disabilities*, 43, 502-513, 2008.

*DIPIPI-HOY, C., JITENDRA, A.K., & KERN, L. Effects of time management instruction on adolescents' ability to self-manage time in a vocational setting. *The Journal of Special Education*, 43, 145-159, 2009.

*DOTTO-FOJUT, K.M., REEVE, K.F., TOWNSEND, D.B., & PROGAR, P.R. Teaching adolescents with autism to describe a problem and request assistance during simulated vocational tasks. *Research in Autism Spectrum Disorders*, 5, 826-833, 2011.

GAST, D.L., & LEDFORD, J.R. (Eds.). *Single case research methodology:* Applications in special education and behavior sciences. 2 Ed. New York, NY: Routledge, 2014.

*GOH, A.E., & BAMBARA, L.M. Video self-modeling: a job skills intervention with individuals with intellectual disability in employment settings. *Education and Training in Autism and Developmental Disabilities*, 48, 103-119, 2013.

GRIGAL, M., & DESCHAMPS, A. Transition education for adolescents with intellectual disabiltiy. In M.L. Wehmeyer & K.W. Webb (Eds.), *Handbook of Adolescent Transition Education for Youth with Disabilities*. New York, NY: Routledge, 2012. pp. 398-416.

HANLEY-MAXWELL, C. & IZZO, M.V. Preparing students for the 21st Century workforce. In M.L. Wehmeyer & K.W. Webb (Eds.), *Handbook of Adolescent Transition Education for Youth with Disabilities*. New York, NY: Routledge, 2012. pp. 139-155

HENDRICKS, D.R., & WEHMAN, P. Transition from school to adulthood for youth with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, 24(2), 77-88, 2009.

HENDRICKS, D.R. Employment and adults with autism spectrum disorders: challenges and strategies for success. *Journal of Vocational Rehabilitation*, 32, 125-134, 2010.

HORNER, R.H., CARR, E.G., HALLE, J., MCGEE, G., ODOM, S., & WOLERY, M. The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71(2), 165-179, 2005.

*HUME, K., & ODOM, S. Effects or an individual work system on the independent functioning of students with autism. *Journal of Autism and Developmental Disorders*, 37, 1166-1180, 2007.
Individuals with Disabilities Education Improvement Act (IDEA) of 2004, PL 108-446, 20 U.S.C. ss 1,400 et seq.

KAZDIN, A. E. *Single-case research designs:* Methods for clinical and applied settings. 2 Ed. New York, NY: Oxford University Press, 2011.

*KELLY, J.A., WILDMAN, B.G., & BERLER, E.S. Small group behavioral training to improve the job interview skills repertoire of mildly retarded adolescents. *Journal of Applied Behavior Analysis*, 13, 461-471, 1980.

*KEMP, D.C., & CARR, E.G. Reduction of severe problem behavior in community employment using an hypothesis-driven multicomponent intervention approach. *Journal of the Association of Persons with Severe Handicaps*, 20(4), 229-247, 1995.

KRATOCHWILL, T.R., HITCHCOCK, J., HORNER, R.H., LEVIN, J.R., ODOM, S.L., RINDSKOPF, D.M., & SHADISH, W.R. *Single-case design technical documentation*. Disponível em: <http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf> 2010.

KRATOCHWILL, T.R., HITCHCOCK, J.H., HORNER, R.H., LEVIN, J.R., ODOM, S.L., RINDSKOPF, D.M., SHADISH, W.R. Single-case intervention research design standards. *Remedial and Special Education*, 34(1), 26-38, 2013.

*LATTIMORE, L.P., PARSONS, M.B., & REID, D.H. Enhancing job-site training of supported workers with autism: A reemphasis on simulation. *Journal of Applied Behavior Analysis*, 39, 91-102, 2006.

*LATTIMORE, L.P., PARSONS, & M.B., REID, D. Simulation training of community job skills for adults with autism: A further analysis. *Behavior Analysis in Practice*, 1, 24-29, 2008.

*LATTIMORE, L.P., PARSONS, & M.B., REID, D. Rapid training of community job skill to nonvocal adults with autism: An extension of intensive training. *Behavior Analysis in Practice*, 2, 34-42, 2009.

*LIKINS, M., SALZBERG, C.L., STOWITSCHEK, J.J., LIGNUGARIS/KRAFT, & CURL, R. Co-worker implemented job training: the use of coincidental training and quality-control checking on the food preparation skills of trainees with mental retardation. *Journal of Applied Behavioral Analysis*, 22, 381-393, 1989.

LIPSEY, M.W., & WILSON, D.B. *Practical meta-analysis.* Applied Social Research Methods Series. v.49. Bickman, L., & Rog, D.J. (Eds.). Thousand Oaks, CA: Sage Publications, 2001.

MAGGIN, D.M., BRIESCH, A.M., & CHAFOULEAS, S.M. An application of the What Works Clearinghouse Standards for evaluating single-subject synthesis of the self-management literature base. *Remedial and Special Education*, 34, 44-58, 2013.

*MARTIN, J.E., ELIAS-BURGER, S., & MITHUAG, D.E. Acquisition and maintenance of time-based task change sequence. *Education and Training in Mental Retardation*, 22(4), 250-255, 1987.

MASON, R.A., DAVIS, H.S., BOLES, M.B., & GOODWYN, F. Efficacy of point-of-view video modeling: A meta-analysis. *Remedial and Special Education*, 34, 333-345, 2013.

MASON, R.A., GANZ, J.B., PARKER, R.I., BURKE, M.D., & CAMARGO, S.P. Moderating factors of video-modeling with other as model: A meta-analysis of single-case studies. *Research in Developmental Disabilities*, 33, 1076-1086, 2012.

*MECHLING, L.C., & ORTEGA-HURNDON, F. Computer-based video instruction to teach young adults with moderate intellectual disabilities to perform multiple step, job tasks in a generalized setting. *Education and Training in Developmental Disabilities*, 42, 24-37, 2007.

*MECHLING, L.C., & SAVIDGE, E.J. Using a personal digital assistant to increase completion of novel tasks and independent transitioning by students with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 41, 687-704, 2011.

*MECHLING, L.C., & AYRES, K.M. A comparative study: completion of fine motor office related tasks by high school students with autism using video models on large and small screen sizes. *Journal of Autism and Developmental Disorders.* 2012.

*MITCHELL, R.J., SCHUSTER, J.W., COLLINS, B.C., & GASSAWAY, L.J. Teaching vocational skills with a faded auditory prompting system. *Education and Training in Mental Retardation and Developmental Disabilities*, 35, 415-427, 2000.

*MORGAN, R.L., & SALZBERG, C.L. Effects of video-assisted training on employment-related social skills of adults with severe mental retardation. *Journal of Applied Behavior Analysis*, 25, 365-383, 1992.

NEWMAN, L., WAGNER, M., CAMETO, R., & KNOKEY, A.-M. *The Post-High School Outcomes of Youth with Disabilities up to 4 Years After High School. A Report of Findings from the National Longitudinal Transition Study-2 (NLTS2) (NCSER 2009-3017).* Menlo Park, CA: SRI International. Disponível em: <*www.nlts2.org/reports/2009_04/nlts2_report_2009_04_complete.pdf*> 2009.

NEWMAN, L., WAGNER, M., KNOKEY, A.-M., MARDER, C., NAGLE, K., SHAVER, D., … SCHWARTING, M. *The post-high school outcomes of young adults with disabilities up to 8 years after high school: A report from the National Longitudinal Transition Study-2 (NLTS2)* (NCSER 2011-3005). Menlo Park, CA: SRI International. Disponível em: <*www.nlts2.org/reports/*> 2011.

NINCI, J., NEELY, L.C., HONG, E.R., BOLES, M.B., GILLILAND, W.D., GANZ, J.B., … VANNEST, K.J. Meta-analysis of single-case research on teaching functional living skills to

individuals with ASD. *Review Journal of Autism and Developmental Disorder*, 2, 184-198, 2015.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, sec. 115, Stat. 1425, 2002.

PALMEN, A., DIDDEN, R., & LANG, R. A systematic review of behavioral intervention research in adaptive skill building in high functioning young adults with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 6, 602-617, 2012.

*PARSON, M.B., REID, D.H., GREEN, C.W., BROWNING, L.B. Reducing individualized job coach assistance provided to persons with multiple severe disabilities in supported work. *Research and Practice for Persons with Severe Disabilities*, 24, 292-297, 1999.

REICHOW, B., VOLKMAR, F.R., CICCHETTI, D.V. Development of the evaluative method for evaluating evidence-based practices in autism. *Journal of Autism and Developmental Disorders*, 38, 1311-1319, 2008.

*RIFFEL, L.A., WEHMEYER, M.L., TURNBULL, A.P., LATTIMORE, J., DAVIES, D., STOCK, S., & FISHER, S. Promoting independent performance of transition-related tasks using a palmtop PC-based self-directed visual and auditory prompting system. *Journal of Special Education Technology*, 20(2), 5-14, 2005.

ROTH, M.E., GILLIS, J.M., & DIGENNARO-REED, F.D. A metaanalysis of behavioral interventions for adolescents and adults with autism spectrum disorders. *Journal of Behavioral Education*, 23, 258-286, 2014.

RUSCH, F.R., & DATTILO, J. Employment and self-management: A meta-evaluation of seven literature reviews. *Intellectual and Developmental Disabilties*, 50(1), 69-75, 2012.

SANFORD, C., NEWMAN, L., WAGNER, M., CAMETO, R., KNOKEY, A.-M., & SHAVER, D. *The Post-High School Outcomes of Young Adults With Disabilities up to 6 Years After High School. Key Findings From the National Longitudinal Transition Study-2 (NLTS2)* (NCSER 2011-3004). Menlo Park, CA: SRI International. Disponível em: <*www.nlts2.org/reports/*> 2011.

SCHWARTZ, I.S., & BAER, D.M. Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis*, 24, 189-204, 1991.

TAYLOR, J.L., MCPHEETERS, M.L., SATHE, N.A., DOVE, D., VEENSTRA-VANDERWEELE, J., & WARREN, Z. *Pediatrics*, 130, 531-538, 2012.

*VAN LAARHOVEN, T., VAN LAARHOVEN-MYERS, T., & ZURITA, L.M. The effectiveness of using a pocket PC as a video modeling and feedback device for individuals with developmental disabilities in vocational settings. *Assistive Technology Outcomes and Benefits*, 4(1), 28-45, 2007.

*WACKER, D.P., BERG, W.K., BERRIE, P., & SWATTA, P. Generalization and maintenance of complex skills by severely handicapped adolescents following picture prompt training. *Journal of Applied Behavior Analysis*, 18, 329-226, 1985.

*WACKER, D.P., BERG, W.K., CHOISSER, L., & SMITH, J. Evaluation of the generalize effects of a peer-training procedure with moderately retarded adolescents. *Journal of Applied Behavior Analysis*, 22, 261-273, 1989.
WALKER, A.R., RICHTER, S., UPHOLD, N.M., TEST, D.W. Review of the literature on community-based instruction across grade-levels. *Education and Training in Autism and Developmental Disabilities*, 45(2), 242-267, 2010.

WEHMEYER, M.L. Employment status and perceptions of control of adults with cognitive and developmental disabilities. *Research in Developmental Disabilities*, 15(2), 119-131, 1994.

WOLERY, M. A commentary: Single-case design technical document of the What Works Clearinghouse. *Remedial and Special Education*, 43, 39-43, 2013.

WOLF, M.M. Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behaivor Analysis*, 11, 203-214, 1978.

MARGOT BOLES, Ph.D., BCBA-D is a Doctoral Graduate at Texas A&M University in College Station, TX. Her areas of research interest are interventions promoting social skills, functional daily living skills, and career skills for children, adolescents, and adults with autism spectrum disorder (ASD) and other developmental disabilities, transition from school to work/postsecondary education, video modeling and educators, therapists, and caregivers training to implement evidence-based interventions in classroom, clinical, and institutional settings. E-mail: margotboles@gmail.com

JENNIFER B. GANZ, Ph.D., BCBA-D is a Professor of Special Education Department of Educational Psychology and Affiliated Faculty, Center on Disability & Development at Texas A&M University in College Station, Texas. Her areas of research interest are development and efficacy of augmentative and alternative communication interventions, development and investigation of interventions to impact social-communication deficits and efficacy of technology- and visually-based interventions to improve communication, social skills, and behavior in individuals with autism spectrum and intellectual disabilities. E-mail: jeniganz@tamu.edu

SHANNA HAGAN-BURKE, Ph.D. is an Associate Professor of Special Education in the Department of Educational Psychology and Associate Department Head at Texas A&M University in College Station, Texas. Her areas of research interest are relations between academic performance and problem behavior, functional analyses of problem behavior, early

literacy and positive behavior interventions and supports (PBIS). E-mail: shaganburke@tamu.edu

EMILY V. GREGORI, M.Ed. is a Doctoral Student at Texas A&M University, College Station, TX. Her areas of research interests are Assessment and treatment of challenging behavior in adults with AU and IDD. E-mail: egregori1@tamu.edu

LESLIE C. NEELY, Ph.D., BCBA-D is an Assistant Professor of School Psychology, Department of Educational Psychology at University of Texas at San Antonio in San Antonio, TX. Her areas of research interests are the application of applied behavior analysis to the functional assessment and treatment of challenging behavior for individuals with autism and developmental disabilities and evaluation of acquisition and sustained use of evidence-based practices by parents, teachers, and other interventionists working with individuals with developmental disabilities. E-mail: Leslie.Neely@utsa.edu

ROSE A. MASON, Ph.D., BCBA-D is an Assistant Research Professor at University of Kansas in Kansas City, Kansas; Juniper Gardens Children's Project. Her areas of research interest are developing and evaluating socially valid interventions for individuals with ASD across the lifespan with particular focus on contextual variables that strengthen the efficacy of the intervention and facilitate research to practice through identification of sustainable methods of training and feedback for practitioners who intervene with children and adolescents with challenging behaviors. E-mail: rmason519@ku.edu

DALUN ZHANG, Ph.D. is a Professor of Special Education Department of Educational Psychology and Director of the Center on Disability & Development at Texas A&M University in College Station, Texas. His areas of research interest are self-determination and transition education and services for individuals with disabilities. E-mail: dalun@tamu.edu

VICTOR WILLSON, Ph.D. is a Professor, Educational Psychology; Department Head, Educational Psychology at Texas A&M University in College Station, Texas. His areas of research interest are statistics, research design, measurement and evaluation, science education, learning and cognition, and reading development. E-mail: v-willson@tamu.edu