# USING NEW CEC QUALITY INDICATORS AND STANDARDS TO DETERMINE THE EVIDENCE BASE OF CLASSWIDE PEER TUTORING

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EDUCATION

MAY 2014

By

Sara Cothren Cook

Dissertation Committee

Bryan Cook, Chairperson
Lysandra Cook
Mary Jo Noonan
Garnett Smith
Katherine Ratliffe

## DEDICATION

I dedicate this to my husband, Cory, and my two daughters, Sarena and Corinne. Cory, I could never have completed this work without out your continued love, support, and encouragement. Thank you for doing so much for our family during my entire doctoral program. Sarena and Corinne, thank you for your unconditional love and the ability to make me laugh even on the toughest days.

## ACKNOWLEDGEMENTS

As I wrap up this part of my life, I feel an overwhelming sense of gratitude for the guidance and support I have received from the faculty in the Department of Special Education at the University of Hawaii at Manoa (UH). I want to acknowledge Dr. Bryan Cook for his continuous support and guidance from the very beginning of my doctoral program. I have so much respect for the time Dr. Cook selflessly puts in with students and cannot thank him enough for the countless hours he spent working with me. Dr. Cook provided me professional advice, encouragement, continuous feedback, and several research opportunities that I would not have been otherwise afforded. My future research agenda was significantly shaped by work I was able to do with Dr. Cook and I am so very grateful.

I want to sincerely thank my dissertation committee for their support and feedback. I want to especially thank Dr. Lysandra Cook for her willingness to spend such a significant amount of personal time coding studies to help me complete my research. I appreciate her patience and the time she spent working with me over the past year.

Thank you to my coworkers and friends who have helped me along this journey. I am so grateful for my former co-teacher and friend, Elizabeth Shiraki. The days we spent teaching together were so valuable to my work at UH and taught me so much about the power of successful collaboration. Dr. Dawn Paresa, thank you for your friendship and encouragement, especially during last few months of the writing process.

Thank you to my parents who have always believed in me. Dad, thank you for coming out to help with Sarena in the early part of my writing. Mom, thank you for your support along the way and your constant encouragement; knowing that I make you so proud motivates me daily.

## TABLE OF CONTENTS

## LIST OF TABLES

## ABSTRACT

In order for students with disabilities to have the opportunity to meet the same academic standards and expectations as their peers in general education, students with disabilities are, more than ever, educated within the general education classroom. Placement in general education will not alone ensure the success of students with disabilities; it is essential that teachers use the most effective instructional strategies. Evidence based practices (EBPs) represent the most recent efforts to identify what works in education. A Council for Exceptional Children Workgroup has recently developed quality indicators and standards for determining EBPs in special education. The purpose of this study was to determine (a) the inter-rater reliability for Cook et al.'s (2013) proposed quality indicators and standards for special education and (b) whether classwide peer tutoring (CWPT) is an EBP for students with mild disabilities. Sixteen single subject studies met inclusion criteria; five were coded for inter-rater reliability. Kappa statistics for individual studies ranged from k = 0.16 to k = 1.0. Combined kappa was 0.64, which suggests substantial agreement (Landis & Koch, 1977). Percentage agreement scores were calculated for individual quality indicators. Inter-rater reliability was perfect (100%) for the majority of quality indicators (13/23), moderate (80%) for five quality indicators, and low (60%) for five quality indicators. None of the 16 studies were considered to be methodologically sound; therefore it was determined that there is currently insufficient evidence for CWPT to be considered an EBP for students with mild disabilities.

# CHAPTER 1: INTRODUCTION

Evidence based practices (EBPs) are instructional practices shown to have positive outcomes for students with disabilities (Kretlow & Blatz, 2011). In order for an intervention to be considered evidenced based, it must be supported by a body of rigorous research. Generally, supporting research must meet criteria related to: (a) design, (b) quality, (c) quantity, and (d) effect size. That is, research must demonstrate evidence of a practice's effectiveness from multiple studies that are considered to be high quality and of appropriate design.

EBPs are considered to be one of the essential components of bridging the research to practice gap in education (e.g., Slavin, 2002). The federal government has recognized the importance of establishing and implementing EBPs in order to increase the academic performance of all students; in 2002, the Institute of Education Sciences (IES) established the What Works Clearinghouse (WWC) to provide educators, policy makers, researchers, and the public with a source of evidence about 'what works' in education" (WWC, 2011). The WWC reviews research in 15 topic areas (e.g., English Language Learners, special education, secondary math) to establish EBPs in education.

The WWC is not the only organization to establish criteria and standards for determining EBPs. In fact, several other organizations have established protocols for determining "what works" in education (e.g., Best Evidence Encyclopedia, Promising Practice Network). However, various organizations have different standards that vary in terms of rigor when establishing interventions as EBPs. Because different organizations have different methods of determining EBPs, an intervention that is established as an EBP by one organization may not be considered an EBP by another organization. For

example, the Best Evidence Encyclopedia established classwide peer tutoring (CWPT) as an EBP for students in elementary math, but the WWC has not reviewed CWPT in elementary math (and found only potentially positive effects for reading). That said, organizations generally follow four basic steps in determining whether a practice is an EBP (WWC, 2011). Reviewers must:

1. Target an intervention and topic area for review (e.g., CWPT for students with disabilities).
2. Locate all studies that use an acceptable research design (e.g., group experimental, quasi-experimental, single subject design) and meet additional inclusion criteria (e.g., critical components of strategy, participant demographics).
3. Code all extant research using pre-determined quality indicators in order to identify only high quality studies for review. Quality indicators refer to the standards of methodological rigor (e.g., fidelity of implementation, measures of generalized performance, attrition rates, description of participants) that are present in high quality, trustworthy research studies. Organizations will include studies in an evidence-based review only if they meet certain standards of methodological rigor.
4. Apply pre-determined criteria to high quality studies to establish whether the intervention meets standards of an EBP. Organizations establish criteria for how many high quality studies showing positive effects are needed to establish an intervention as an EBP. Magnitude of effect size may also be considered.

**Statement of the Problem**

Different organizations use different quality indicators and standards, but they also review different topic areas and interventions and contain little direction for special education teachers. For example, the Promising Practice Network (2013) and Best Evidence Encyclopedia (n.d.) review interventions for 29 and 15 different topic areas, respectively; but neither organization includes reviews specifically for students in special education. And although the WWC does have topic areas targeted for students with disabilities (i.e., Children and Youth with disabilities, Emotional Behavior Disabilities (ED), Learning Disabilities (LD)), the organization tends to focus their reviews on broad educational programs (e.g., Alphabet Phonics, Barton Reading & Spelling Systems) rather than discrete interventions (e.g., CWPT, repeated reading).

In special education, the focus of establishing EBPs has been more on the level of discrete interventions (Cook, Tankersley, & Landrum, 2009). When implementing discrete interventions, special educators have more flexibility in adapting the intervention to meet the needs of students with disabilities than when implementing comprehensive programs or curricula (Cook & Cook, 2011); this is important in special education because of the complexity and variability of student needs, which necessitates that teachers be able to adjust their instruction to meet the needs of particular students. Therefore, when it comes to identifying EBPs in special education, researchers cannot only focus on whether broad programs are effective, but should also consider whether practices that can be applied for small groups and individual students are effective. This individualized focus of instruction in special education further necessitates a focus on *who* practices are effective for (e.g., students with LD, students with autism) (Guralnick,

1999). Because organizations reviewing EBPs infrequently: (a) focus on interventions specifically for students with disabilities or (b) focus on reviewing discrete interventions that allow flexibility on the part of the special education teacher, it is important for the field of special education to conduct their own EBP reviews and establish EBPs specifically for students with disabilities.

CWPT is an instructional strategy that is based on reciprocal peer tutoring and group reinforcement. CWPT was developed in 1980 and has since been used as an instructional strategy for students with and without disabilities (e.g., Arreaga-Mayer, 1998). As previously mentioned, several general education organizations have reviewed CWPT using their evidence based standards (e.g., WWC, Best Evidence Encyclopedia) with mixed results. And although there has been 30 years of research on the topic of CWPT, these organizations do not review the research to determine what specific populations of students with disabilities may benefit from CWPT. It will be important for special education researchers to review CWPT research for students with disabilities. As Cook and Schirmer (2006) noted, although students without disabilities may achieve success without the use of effective instructional strategies, it is absolutely necessary for students with disabilities to be taught with highly effective strategies in order to make academic gains.

To date, the Council for Exceptional Children (CEC), the largest professional organization devoted to the education of exceptional children, has not officially adopted a set of quality indicators and standards for determining EBPs in special education. However, groups of special education researchers have proposed quality indicators and standards for determining EBPs specifically for students with disabilities (i.e., Gersten et

al., 2005; Horner et al., 2005). In 2005, Gersten et al. and Horner et al. defined and described quality indicators and standards for determining EBPs for group experimental design and single subject design, respectively. After developing the 2005 quality indicators and standards, Gersten et al. and Horner et al. encouraged special education researchers to the field test these proposed standards in order to refine and determine a set of quality indicators and standards to be used by special education researchers.

In a 2009 special issue of *Exceptional Children* (Cook, Tankersley, & Landrum), five review teams of special education researchers field tested the 2005 proposed quality indicators and standards for group experimental design and single subject design. The review teams made suggestions regarding how to refine and improve quality indicators and standards. Results of the reviews indicated the need to operationalize and clearly define the process of determining EBPs in special education. Specifically, Cook et al. (2009) noted how the review teams interpreted and determined the presence of the quality indicators very differently. Cook et al. (2009) explained that Gersten et al. (2005) and Horner et al. (2005) were asked to identify and briefly explain quality indicators and standards, but not to operationally define them; therefore, researchers in special education are left to their own interpretation of the quality indicators. For example, whereas some researchers rated the presence of quality indicators on a dichotomous scale (i.e., met or not met), others used 4-point rubrics to determine the presence of indicators. In order to formalize the process for identifying EBPs in special education, standards for EBPs, quality indicators for individual studies, and how both sets of criteria are to be measured must be operationally defined and agreed upon.

CEC has appointed a work group charged with developing, approving, and piloting new standards for EBPs in special education (Cook et al., 2013). Cook and colleagues expanded on the work of Gersten et al. (2005) and Horner et al. (2005) by using (a) the 2005 indicators proposed by Gersten et al. and Horner et al., (b) feedback and suggestions from the 2009 field testing in order to improve quality indicators and standards for EBPs, and (c) input on a draft version of the revised standards from 23 expert researchers in special education who participated in a Delphi study. From this information, Cook et al. (2013) have proposed new quality indicators and standards for group experimental designs and single subject designs. However, because the quality indicators and standards are newly developed, they have yet to be field tested. Field testing is the first step in determining whether the 2013 quality indicators and standards are clearly defined and operationalized.

**Purpose of Study**

The primary purpose of this study is to assess inter-rater reliability of the 2013 quality indicators and standards. Specifically, I coded studies investigating the effects of CWPT using the 2013 quality indicators proposed by Cook et al. (2013) for group comparison and single subject research designs. Secondly, I calculated inter-rater reliability scores for the 2013 quality indicators to determine whether the new standards are clearly defined and operationalized.

The second purpose of this study is to determine whether CWPT can be considered an EBP for students with learning disabilities using the 2013 standards for special education. Specifically, I used the 2013 standards for determining EBPs proposed

by Cook et al. (2013) for group comparison research and single subject research to determine whether CWPT can be considered an EBP.

**Rationale.** The rationale for this study is based on the following premises:

1. The No Child Left Behind Act (NCLB) of 2001 and Individuals with Disabilities Education Act (IDEA) (2004) had significant impact on the expectations for students with disabilities in school. More than ever, students with disabilities are being held to the same standards and expectations as their peers in general education. Specifically, more students with disabilities are being educated in the general education classroom. However, placement in the general education does not ensure the success of students with disabilities in meeting grade level standards. Therefore, both NCLB and IDEA require teachers choose interventions and strategies supported by scientifically based research.
2. EBPs represent researchers' most recent efforts to identify what works in education (Cook & Cook, 2011). Specifically, EBPs are instructional strategies and techniques that are supported by a "trustworthy body of research that meets specific standards or rigor" (Cook & Cook, 2011, p. 2). EBPs are considered to be an essential tool in bridging the research to practice gap (e.g., Slavin, 2002). To determine whether a strategy is an EBP, a researcher must: (a) identify an instructional strategy to review, (b) find all studies that meet the inclusion criteria (e.g., critical components of strategy, participant demographics); (c) use quality indicators to determine high quality

studies; and (d) code high quality studies against EBP standards to determine whether the instructional strategy is an EBP.

3. Many researchers have documented the research to practice gap (i.e., the problem with educators not using scientifically based research in their classroom) (e.g., Carnine, 1995, 1997; Deschler, 2003; Landrum, Cook, Tankersley, & Fitzgerald, 2002; Odom, 2008; Slavin, 2002; Wanzek & Vaughn, 2006). Additionally, researchers have often speculated on why the research to practice gap exists; for example, (a) teachers find colleagues more than trustworthy than research found in journals (Landrum et al., 2002), (b) research is inaccessible to teachers (Carnine, 1997), and (c) researchers are unable show causal connections in studies published in professional journals (Kennedy, 1997). Researchers have also speculated on ways to bridge the gap: (a) determine teacher perspectives and identify effective and ineffective practices (Burns & Ysseldyke, 2009); (b) use implementation science and enlightened professional development (Odom, 2008); and (c) increase collaboration among researchers and practitioners (Wanzek & Vaughn, 2006). Determining EBPs in special education will not, in itself, bridge the research to practice gap; however, identifying what works in special education is one of the necessary steps for practitioners using instructional strategies that work (Cook & Cook, 2011).
4. Many special education practices documented as effective and research based may be too time consuming to be used for all students (Cook & Schirmer, 2006) and therefore may never be reviewed by non-special education

organizations (e.g., WWC). Therefore, it is important for the field of special education to establish quality indicators and standards for establishing EBPs for students with disabilities. Doing so is especially important for students with disabilities. As Cook and Schirmer (2006) noted, "whereas learners without disabilities will likely succeed without these instructional techniques [EBPs], learners with disabilities fail without them" (p. 179). In order to promote academic success for students with disabilities, researchers in special education need to determine EBPs for students with disabilities, instead of simply relying on what EBPs exist for students in general education.

5. Whereas many organizations have adopted quality indicators, standards, and processes for determining EBPs (e.g., the WWC), CEC has not formally adopted a process for determining EBPs for students with disabilities. Gersten et al. (2005) and Horner et al. (2005) identified and described sets of quality indicators and standards for group experimental and single subject research in special education, respectively. In 2009, teams of special education researchers piloted these indicators (see Browder, Ahlgrim-Delzell-Spooner, Mims, & Baker, 2009; Chard, Ketterlin-Geller, Baker, Doabler, & Apichatabutra, 2009; Lane, Kalberg, & Shepcaro, 2009; Montague & Dietz, 2009) and provided feedback on how the process for determining EBPs in special education could be improved. For example, Baker et al. suggested that although the 2005 quality indicators and standards indicated what should be included in high quality studies, the 2005 standards do not specify how to measure each quality indicator.

Most recently, Cook et al. (2013) proposed revised quality indicators and standards based off the feedback from the research reviews in 2009. Specifically, the 2005 quality indicators needed to be operationalized. In addition, the 2013 standards proposed by Cook et al. proposed how a combination of group experimental and single subject research studies can determine an EBP; this was not a consideration in the 2005 standards.

The 2013 quality indicators and standards proposed by Cook et al. (2013) have yet to be field tested. In order for CEC to adopt a set of quality indicators and standards to be used in special education, researchers must (a) determine if the quality indicators are defined (i.e., operationalized) well enough for different researchers to review studies with high inter-rater reliability, (b) agree that the quality indicators encompass what defines high quality research, and (c) determine that the proposed standards encompass the appropriate number of studies appropriate for determining EBPs.

The revised set of quality indicators and standards for determining EBPs in special education were created and vetted by 23 expert special educators through a Delphi study. However, in order to adopt standards that are meaningful, it is important to ensure proposed standards can be applied reliably (e.g., researchers can apply standards with high inter-rater reliability scores). The next step in establishing EBP standards is to field test the 2013 quality indicators and standards to determine inter-rater reliability. Detailed descriptions of the proposed quality indicators and standards proposed will be outlined in Chapter 2.

6. CWPT has been referred to as an effective intervention for academic engagement and achievement for students with mild or high incidence disabilities (e.g., Greenwood, 1997) in several content areas. Additionally, CWPT has over 20 years of research that supports its implementation for students with disabilities (e.g., Burks, 2004; Delquadri, Greenwood, Stretton & Hall, 1983; Kamps, Barbetta, Leonard, Delquadri, 1994; Maheady & Harper, 1987). However, CWPT has not been reviewed or established as an EBP in special education.

**Importance of the Study and Contribution to Knowledge**

Previous efforts by Gersten et al. (2005) and Horner et al. (2005) provided the field of special education with the first set of quality indicators and standards for determining EBPs for students with disabilities. Since the preliminary publications of this work, several special education researchers have participated in field testing the 2005 quality indicators. Commenting on these field tests, Cook et al. (2009) maintained that, often in these reviews, researchers were left to their own interpretation of the quality indicators. Cook et al. (2009) reported that the original indicators proposed in 2005 by Gersten et al. (2005) and Horner et al. (2005) were not intended to be operationalized and defined, only identified and described. Gersten et al. (2005) emphasized this point and called for the proposed quality indicators and standards to be field tested and refined before being adopted. Cook et al. (2009) emphasized the need for revision and refinement in order to clearly define and standardize the process of determining EBPs in special education.

Cook et al. (2013) reviewed and refined the quality indicators and standards for group design research and single subject research. This study will contribute to the field of special education in the following ways:

1. The study will be the first to field test the quality indicators and standards for identifying EBPs recently proposed by the CEC-appointed task force.
2. The study will examine the inter-rater reliability of the 2013 quality indicators.
3. The study will determine whether CWPT research can be considered an EBP for students with mild disabilities (i.e., learning disabilities, emotional disabilities, ADHD, or mild intellectual disabilities) using the 2013 proposed standards for

EBPs in special education. Although other organizations have determined CWPT as an EBP for students generally (e.g., Best Evidence Encyclopedia), this study will be the first evidence-based review of CWPT specific to students with mild disabilities.

**Research Questions**

1. What are estimates of inter-rater reliability for sets of quality indicators used to identify EBPs in special education?
    a. What is the inter-rater reliability for quality indicators proposed by Cook et al. (2013) for group comparison research across reviewed group comparison studies examining the effects of CWPT on students with mild disabilities?
    b. What is the inter-rater reliability for quality indicators proposed by Cook et al. (2013) for single subject research across reviewed single subject research studies examining the effects of CWPT on students with mild disabilities?
2. Does CWPT meet 2013 standards for an EBP in special education for students with mild disabilities according to standards for identifying EBPs in special education?

**Definition of Terms**

**Classwide peer tutoring.** CWPT is an intervention based on reciprocal peer tutoring and group reinforcement that requires pairs of students to serve as both the tutor and tutee. The goal of CWPT is to facilitate mastery of classroom content and uses a game format (Terry, n.d.). In this study, in order to be considered CWPT, students must:

(a) be placed in partners (pairs), (b) take turns asking and answering questions provided by teacher, (c) be given a set amount of time for each student to be asked questions, and (d) keep track of points earned for correct answers. Additionally, in CWPT, teachers must record: (a) individual points, (b) team points, and (d) daily and/or weekly winners.

**Evidence based practices (EBPs).** EBPs are instructional practices that have been shown by reliable research to have positive outcomes for students with disabilities (Cook & Cook, 2011). In order for an intervention to be considered an EBP the research that supports it must adhere to a specific set of prescribed criteria. Specifically, it must meet criteria in: (a) design, (b) quality, (c) quantity, and (d) effect size.

**EBP standards.** EBP standards refer to the pre-determined quality indicators that define high quality studies of certain design (i.e., random controlled trials, quasi-experimental, single subject) *and* the quantity of high quality studies required when determining whether a strategy can be considered an EBP. Magnitude of effect may also be considered (Cook & Cook, 2011).

**Group comparison research.** Group comparison research designs involve a researcher actively implementing an intervention (e.g., CWPT) on a group of individuals (i.e., treatment group) but not others (i.e., a control group) to see whether there is a change in the dependent variable (e.g., math scores). By showing that implementation of an intervention (i.e., independent variable) changes the dependent variable in the desired direction (e.g., increase in math scores) and the change is meaningful compared to the control group, the researcher can assume that the independent variable caused the change in the dependent variable (as long as the research study was of high methodological quality) (Cook et al., 2008). Group comparison research includes randomized control trials, quasi-experimental research, and regression discontinuity design.

**Quality indicators.** Quality indicators refer to the standards of methodological rigor (e.g., fidelity of implementation, measures of generalized performance, attrition rates, description of participants) used in determining whether a study is high quality, which is a prerequisite for a study to be considered in support of an EBP.

**Quasi-experimental research.** In quasi-experimental designs, researchers use groups that are already in place (Kennedy, 2005). The researchers assign one or more intact groups to be the treatment group and one or more intact groups to be the control. Although using intact groups introduces a potential selection bias, researchers can improve the internal validity of their research by measuring the characteristics between groups and statistically adjusting for identified differences between groups that relate to the outcomes (Cook et al., 2008).

**Single subject research.** When using single subject design, researchers are determining whether there is a functional relationship between the independent variable (i.e., intervention) and dependent variable (i.e., student outcome). Unlike group experimental design, single subject research does not require the use of a control group. In fact one person or one group can function as both the treatment and control. A functional relationship is established when the researcher shows that by introducing the independent variable there is a visible change in the dependent variable. When establishing a functional relationship, the researcher must demonstrate that it is the systematic implementation of the independent variable that is causing the change in the dependent variable, an individual or group may serve as their own comparison. For example, in an ABAB reversal design, a researcher would introduce the independent

variable after baseline. Once the data indicates a change in the dependent variable, the researcher will remove the independent variable. If dependent variable consistently returns to baseline scores after the removal of the independent variable, the researcher has established a functional relationship exists. Possible single subjects designs include (but are not limited to): (a) ABAB design, (b) multi-element design, (c) multiple baseline design, and (d) combined designs (Kennedy, 2005).

**Students with mild disabilities.** Students with mild disabilities are diagnosed with a learning disability, emotional/behavioral disability, attention deficit hyper activity (ADHD), and/or an intellectual disability. Additionally, in some special education research, researchers referred to students with less severe disabilities as mildly disabled, mildly handicapped, or students with mild mental retardation. For the purpose of this research, all of these categories will be referred to as students with mild disabilities.

**Randomized control trials.** Randomized control trials or true experiments are the "gold standard" of special education researcher (WWC, 2003). When researchers are able to use random assignment, they control for extraneous variables (Odom, 2008) and prevent selection bias (Cook et al., 2008). In a true experiment, researchers randomly assign participants to either the treatment or control group. Although this does not ensure that each group is exactly the same, by randomly assigning participants researchers can be sure that they are not contributing to group differences through researcher bias. The goal of researchers in group experimental design is to control for as many extraneous variables as possible in order to maximize the likelihood that the independent variable caused the change in the dependent variable.

# CHAPTER 2: LITERATURE REVIEW

To meet the high demands of the general education curriculum, students with disabilities will need the most effective practices. There seems, however, to be confusion among practitioners as to what really works. Research represents a trustworthy way to analyze whether a practice or intervention can improve outcomes of students with and without disabilities, yet research-validated practices in special education are not consistently implemented in classrooms (Cook & Schirmer, 2006). Additionally, some research studies are more valid than others. EBPs refer to practices supported as effective by studies that adhere to certain standards of methodological quality. Specifically, studies supporting EBPs must meet criteria related to: (a) design, (b) quality, (c) quantity, and (d) effect size. Although many reasons why the research to practice gap exists in special education, slow progress in identifying EBPs for students with disabilities is one contributing factor.

CWPT is an instructional strategy that has been used and researched in special education for over 30 years (Terry, n.d.). Special education researchers have reported that CWPT can result in improved academic outcomes for students with and without disabilities (e.g., Kamps, Barbetta, Leonard, & Delquadri, 1994; Mastropieri, Scruggs, Spencer, & Fontana, 2003). However, CWPT has not been reviewed as an EBP for students with disabilities. In other words, researchers have not conducted a systematic review of the CWPT research literature using standards for EBPs to determine whether a body of trustworthy studies supports CWPT as being effective for students with disabilities.

The purpose of this chapter is to review and summarize the literature base on EBPs and CWPT. This literature review is divided into two major sections: (a) EBPs and (b) CWPT.

## Evidence Based Practices

In the field of special education, EBPs are defined as instructional strategies that have been shown by a number of high quality research studies to positively impact outcomes for students with disabilities (Cook & Cook, 2013). However, the concept of EBPs did not originate from the education field and research on EBPs is prevalent in fields outside of education. In the following sections, I will review: (a) the origin of EBPs; (b) the importance of EBPs in bridging the research to practice gap in education; and (c) the issues in determining EBPs in special education.

**Origin of EBPs.** Sackett (1997) reported the modern EBP movement originated in the field of medicine and can be traced back in history to mid-19th century Paris and earlier. During that time, Pierre Charles Alexandre Louis introduced statistical analysis to evaluate the medical treatment of blood letting. Louis found that this practice had no practical value, but it took many years to bridge the gap between his research and the practice of medical practitioners (Weatherhall, 1996).

However, the *idea* of using clinical trials to make informed decisions about patient care actually dates back even farther:

> In the 17th century Jan Baptista van Helmont, a physician and philosopher, became skeptical of the practice of blood letting. Hence he proposed what was almost certainly the first clinical trial involving large numbers, randomisation and statistical analysis. This involved taking 200 to 500 poor people, dividing them into two groups by casting lots, and protecting one from phlebotomy while

> allowing the other to be treated with as much blood-letting as his colleagues thought appropriate. The number of funerals in each group would be used to assess the efficacy of blood letting. History does not record why this splendid experiment was never carried out (Weatherhall, 1996, p. xi).

When using research to make informed decisions, it is important that practitioners do not rely solely on one research study. By examining the outcomes of several studies, a practitioner may discover that the studies offer conflicting results. When using research to guide practice, medical practitioners rely on research to answer questions regarding a patient's treatment. For example, the Cochrane Collaboration is an international organization whose goal is to help make professionals well informed about health care. Specifically, this group prepares, maintains and promotes access to systematic reviews of evidence in the area of healthcare research in order to provide healthcare providers, consumers, researchers and policy makers with evidence of what works in the field (Higgens & Green, 2011).

The EBP movement is "an effort to ensure scientific knowledge informs the practitioner's decisions regarding intervention" (Detrich, 2008, p. 3-4). In other words, the EBP movement emphasizes the importance of using scientific evidence in combination with professional expertise to make a more informed decision when implementing a treatment or intervention. Evidence based medicine (EBM) refers to the EBP movement in medicine; specifically referring to the practice of using an individual's clinical expertise combined with the best available evidence to make informed decision for a patient's treatment (Sackett, 1997). When seeking scientific evidence to support patients, Sackett suggests practitioners engage in the following activities:

1. Convert information needs into answerable questions.
2. Track down**,** with maximum efficiency, the best evidence with which to answer them.
3. Critically appraise that evidence for its validity and usefulness.
4. Integrate the appraisal with clinical expertise and apply the results in clinical practice; and
5. Evaluate one's own performance (p.4).

Weatherhall (1996) emphasized that EBM should influence everything a medical practitioner does. In other words, finding and utilizing research in combination with their clinical expertise should guide their decisions in recommending treatment for patients. Sackett (1997) clarified the types of research that practitioners should refer to; EBM is not restricted to using studies with randomized trials and meta-analysis.

David Sackett has been credited for advancing the importance of EBP in the medical field in the 20th century (Weatherhall, 1996) and has encouraged more medical practitioners to engage in EBM. However, his work has not been met with "open arms" (Sackett, 1997). One misconception among some practitioners is the idea that EBM replaces the need for clinical expertise and will turn the medical profession into "cookbook" medicine. However, Sackett suggested that this could not be further from the truth:

> External clinical evidence can inform, but can never replace individual clinical expertise, and it is this expertise that decides whether the external evidence applies to the individual patient at all and, if so, how it should be integrated into a clinical decision (Sackett, 1997, p.4).

Although the EBP movement in medicine has advanced over the last 100 years, bringing research to bear on practice continues to be problematic (Greenwood & Abbott, 2001). To illustrate the personal consequences of the research to practice gap in medicine, Greenwood and Abbott described the following real-life scenario regarding the cause and treatment of stomach ulcers:

> Long thought caused by stress, the discovery that bacteria caused stomach ulcers is one of the most amazing medical breakthroughs of this generation. Australian physicians Roving Warren and Barry Marshall provided hard evidence that the bacteria (i.e., Helicobacter pylori) was the cause and not stress in 1982. Today antibiotic medication is the treatment of choice for stomach ulcers, and we know that ulcers can be cured (Centers for Disease Control, 2000).
>
> Yet for Arlene Ozburn suffering ulcer symptoms were slow coming (Kansas City Star, Sunday, Feb 4, 1996). It was two years of pain and stress reduction treatment before a friend suggested that she have her doctor look into this discovery, before her doctor finally checked her for the ulcer bacteria and provided her antibiotic medication (p. 278)!

Over the past decades, the EBP movement has moved to other disciplines (Detrich, 2008) such as farming, nursing, psychology, and education (Slavin, 2002). In the following section, I discuss the EBP practice movement in education, with a specific focus on the development of EBPs in special education.

**EBPs in education.** Like the EBP movement in the field of medicine, the EBP movement in education values the importance of combining scientific evidence with professional judgment to make informed decisions about interventions and practices.

Also similar to the field of medicine, the field of education is also experiencing difficulty in bringing new research findings into practice when it comes to educational strategies and interventions (Greenwood & Abbott, 2001). In the following section, I describe the research to practice gap in education and discuss how the development of EBPs is a primary step in bridging this gap.

***Gap in research to practice.*** The research to practice gap in education refers to "the mismatch between research findings and classroom-level implementation of the practices associated with these findings" (Wanzek & Vaughn, 2006, p. 165). More specifically, a research-to-practice gap exists when educators use practices shown by research to be effective less frequently than practices without research support. Many educational researchers have documented the research to practice gap (e.g., Carnine, 1997; Deshler, 2003; Landrum, Cook, Tankersley, & Fitzgerald, 2002; Odom, 2008; Slavin, 2002, Wanzek & Vaughn, 2006). The literature suggests that both researchers and practitioners contribute to special education's research to practice gap. In order to illustrate what the research to practice gaps "looks like" in education, I first explain how districts, schools and classroom teachers have historically made educational programming decisions. Next, I explain other major factors that contribute the research to practice gap; and last, I explain how developing EBPs is the first step in bridging the research to practice gap in special education.

*School/district and classroom level decisions.* There are two major levels of educational program decision making: (a) district/school level and (b) classroom decisions. Whereas districts and/or schools often choose school wide programs, classroom teachers often choose strategies that they implement in their individual

classrooms. In this section, I (a) describe how these programs and strategies are chosen at the district/school and classroom level and (b) describe how these decisions contribute to the research to practice gap.

District superintendents or school principals often adopt educational programs at the district or school level. Programs decisions made at the school or district level typically involve broad educational programs. Broad educational programs usually take the place of any curriculum already in place and are packaged to include: (a) specific instructional materials, (b) instructional approaches, (c) lessons, and (d) the type and amount of training needed to implement (Cook & Cook, 2011). The "University of Chicago School Mathematics Program: Project 6-12 Curriculum" (WWC, 2011), is an example of a broad educational program. If this program was chosen at the district or school level, all teachers in the school would be expected to implement the yearlong curriculum according to the guidelines set by the University of Chicago (WWC). Oftentimes, school wide programs require teachers to follow specific protocols without a lot of room for adaptation. Districts and schools typically choose these broad educational programs in order to improve academic achievement.

Slavin (1989) reported that districts and schools often choose programs without a strong research base (i.e., without strong evidence of program effectiveness) and instead choose programs based on popularity. In fact, Slavin used a metaphorical pendulum to describe how programmatic decisions and changes occur at the district and school levels. Specifically, in the pendulum's upswing, schools: (a) choose educational programs based on popularity (e.g., a publication of a new idea); (b) pilot the program using flawed data

collection; (c) introduce the program; (d) expand the program rapidly; and (e) begin to evaluate the program's effectiveness.

The downward swing of the pendulum begins when districts and schools begin to receive the preliminary program evaluations. Oftentimes these evaluations are disappointing and lead to program developers claiming that the program was poorly implemented (Slavin, 1989). However, over time, researchers conduct more evaluations that lead to further evidence of program ineffectiveness. In other words, rather than choosing educational programs with research support, districts and schools are implementing programs before researchers have time to conduct high quality research to determine effectiveness. The pendulum's upswing begins again when another "popular" program is introduced (Slavin, 1989). Making educational program decisions before researchers can report on a program's effectiveness contributes to the research to practice gap.

Teachers make daily decisions regarding instructional strategies and interventions to use within a given day or class period. The instructional strategies and interventions teachers choose typically are different from the programs chosen by districts and schools; these strategies do not usually constitute an entire curriculum and allow flexibility in implementation (Cook & Cook, 2011). Examples include CWPT, direct instruction, graphic organizers, and repeated reading strategies. It is important for teachers to choose strategies that have been empirically validated to support students' academic achievement. However, Huberman (1983) explained that teachers often make decisions based on (a) their own intuition, and (b) a focus on short-term outcomes. Perhaps because

of this, teachers may be more likely to use a practice that a fellow colleague uses than one supported by high quality research (Cook, Tankersley, Cook, & Landrum, 2008).

Historically, "the adoption of instructional programs and practices has been driven more by ideology, faddism, politics, and marketing than by evidence" (Slavin, 2008, p. 5). In order for both districts, schools and teachers to use "what works," they must: (a) value the findings of research; (b) have the ability and time to locate interventions and interpret research findings; and (c) have meaningful professional development in order to implement strategies with fidelity (Carnine, 1997). In the following sections, I outline four elements underlying the research to practice gap in special education: (a) the separateness of the research and practice communities; (b) limited relevance of educational research to practitioners; (c) the lack of applicability of research to practitioners, and (d) the lack of effective professional development opportunities involving both practitioners and researchers (Greenwood & Abbot, 2001).

*Separation of research and practice communities.* Most special education researchers and practitioners have the same end goal: to improve the academic and social outcomes of students with disabilities. However, it seems rare that these two communities work together to reach the end goal. Smith, Schmidt, Edelen-Smith, and Cook (2013) suggest that whereas researchers value interventions that are supported by numerous methodologically sound research studies (i.e., EBPs), practitioners value a different approach in identifying what works. Specifically, practitioners prefer identifying what works based on their own values, experiences, and action research (i.e., practice based evidence (PBE). In other words, practitioners focus on external validity (i.e., what works for their students) and are less concerned with the internal validity of the research. In

contrast, a primary goal of educational research is to validate interventions by examining research that is internally valid and rigorous (Smith et al.). The stark contrast in values between researchers and practitioners contributes to the research to practice gap.

Wanzek and Vaughn (2006) suggest that it is during the initial implementation of a new strategy when teachers decide whether or not to continue its use; teachers who understand the implementation of a practice and its conceptual foundations are more likely to use the practice in their classroom (Gersten & Dimino, 2001; Klingner, Ahwee, Pilonieta, & Menendez, 2003, as cited in Wanzek & Vaughn, 2006). Thus, it seems imperative that the research and practice communities work together in order to successfully plan and implement research based practices. In order to help bridge the research to practice gap, it will be important for researchers to support teachers in both understanding the conceptual underpinnings of an intervention and how to implement the intervention with fidelity.

*Limited relevance of educational research.* Greenwood and Abbott (2001) suggested that many practitioners perceive educational research to (a) be inaccurate and (b) have limited relevance to real classroom situations (see also Boardman et al., 2005). Carnine (1997) argued that practitioners have valid concerns about the relevance of educational research; practitioners are and should be concerned with the "trustworthiness" and "usability" of educational research.

Trustworthiness refers to how confident practitioners can be in research findings (Carnine, 1997). In order for practitioners to use interventions recommended in educational research, researchers must convince practitioners that research findings are meaningful and accurate. Therefore, establishing EBPs seems essential. EBPs refer to

instructional practices that have been shown by multiple high quality studies to have meaningfully positive effects on student outcomes (Cook & Cook, 2011). By establishing EBPs, practitioners may become more confident that an intervention will have the impact that the research studies suggest. Conducting numerous high quality studies that show a strategy's effectiveness will also help alleviate practitioners' concerns of contradictory research findings (Flemming, 1988, as cited in Greenwood & Abbott, 2001).

Another issue related to trustworthiness is the confusion among terms such as "best practice," "research based practice," and "evidence based practice." Terms such as "best practice" and "research based practice" have been used at countless professional development workshops for teachers, yet not all of these practices may really be effective in supporting students with disabilities (Cook & Cook, 2011). For example, Cook and Cook noted that "evidence-based practice" is often used inappropriately, referring to practices supported by some research (e.g., a single, low quality study), and that "best practices" and "research based practices" can refer to practices with little or no actual research support. Consequently, teachers who choose to implement strategies labeled as "best practice" or "researched based practice" may find the practice does not produce desired results and begin to distrust educational research findings. Thus, it is important for the educational community to define what is meant by terms such as EBP and to use terms appropriately.

The usability of research also contributes to the practitioner perspective that educational research is often irrelevant. Tinkunoff and Ward (1983) reported that teachers "frequently have been given answers to questions they never asked and solutions to problems they never had" (p. 454). "Usability" refers to the likelihood that the education

research on a practice is used by those who actually teach students (Carnine, 1997). Although Carnine commended the special education community for the usability of its research, it is important to note that: (a) in general, educational research is not seen as useable; and (b) even when research is useable, it is not often accessible to practitioners.

*Difficulty translating research faithfully into practice.* Another major contribution to the research to practice gap is the many obstacles in translating research findings to classrooms (Greene & Abbott, 2001). In other words, it has been historically difficult for practitioners to implement practices found to be successful in research with high fidelity. Carnine (1997) suggested that if practitioners find it difficult to locate and interpret research, it is likely they will not make an effort to use it. Therefore, although it is important that practitioners are able to quickly and easily obtain and implement research findings, researchers must ensure that research findings provide enough information in order to practitioners to be confident in implementation. If researchers do not provide enough information for replication, practitioners may (a) implement with low fidelity (resulting in diminished effects) (Stallings, 1975; as cited in Greenwood & Abbott, 2001) or (b) not attempt to use the practice at all.

*Lack of meaningful professional development.* Greenwood and Abbot (2001) cited the lack of meaningful professional development as a contributor to the research to practice gap. Professional development opportunities often occur over a brief period of time (e.g., one or two days, or even hours). Odom (2008) referred to one shot workshops as "expired," meaning that these types of professional development are no longer relevant and meaningful in bridging the research to practice gap. In fact, research has shown that

it takes more than 50 hours of professional development for practitioners to develop and use a new skill (Darling-Hammond, Wei, Richardson, & Orphanos, 2009).

One shot workshops traditionally use a "top down" approach, meaning practitioners are told how to teach. Garet, Porter, Desimone, Birman, and Yoon (2001) indicated that when professional development was integrated into the daily school life it was more effective. Enlightened professional development approaches refer to approaches to professional development that go beyond the one shot, top down workshop (e.g., coaching and consultation, communities of practice, online instruction) (Odom, 2008) and are intended to enhance the use of EBPs in the field of education. Enlightened professional development emphasizes the need of integrating researcher and practitioner communities.

Just as there are many contributions to research to practice gap in education, there are multiple aspects involved in bridging this gap. I briefly discussed the need for researchers to: (a) research practices that practitioners will find useable, (b) establish EBPs, (c) support practitioners in initial implementation of EBPs, and (d) provide enlightened professional development to ensure implementation with fidelity. In the next section of this literature review, I discuss the field of education's progress in determining EBPs. I then explain: (a) the differences in general and special education relevant to EBPs and (b) the importance of special education researchers to adopt their own standards for establishing EBPs.

**Establishing EBPs in general education.** The What Works Clearinghouse (WWC) was established in 2002 by the U.S. Department of Special Education's Institute of Education Sciences (IES). The WWC has since reviewed thousands of studies on

different educational programs, products, practices, and policies to establish EBPs in the field of education. For each review, the WWC review teams:

1. Create review protocols in order to establish inclusion and exclusion criteria including: (a) how studies will be identified (i.e., search procedures), (b) outcomes that will be identified (e.g., academic outcomes), (c) time period for studies, and (d) key words for search.
2. Identify relevant studies through systematic literature search.
3. Review studies for relevance to the topic and adequacy of study design, implementation, and reporting.
4. Gather and summarize information on the program, practice, product, or policy studied, the study's characteristics and the study's findings.
5. Combine findings within and across settings in order to determine the effectiveness of the program, practice, product, or policy.

The WWC currently uses this five step process to review studies in 15 different topic areas (e.g., academic achievement, dropout prevention, English Language Learners, literacy, math, student behavior, and special needs). After the review team completes a comprehensive review on a particular program, practice, product, or policy, the intervention is rated as having positive, potentially positive, mixed, no discernable, potentially negative, or negative effects (WWC, 2008).

Currently, the WWC (2011) considers only randomized controlled trials and quasi-experimental designs for determining the effectiveness of an intervention. When reviewing the quality of the research design, the WWC rates studies as meeting evidence standards, meeting evidence standards with reservations, or not meeting evidence

standards. Only randomized controlled trials can meet evidence standards without reservations. Studies that use quasi-experimental research design can only meet evidence standards with reservations. Currently, the WWC has established pilot criteria for single subject research (SSR) and regression discontinuity (RD) designs; yet SSR and RD studies are not currently reviewed by the WWC.

The WWC (2011) has established clearly defined standards in order to determine the methodological quality of reviewed studies (e.g., low overall and differential attrition, no evidence of intervention contamination). After studies are reviewed for methodological quality, review teams establish the overall effectiveness of an intervention. Review teams use studies' effect size and statistical significance to rate the effectiveness of a particular intervention (e.g., positive effects, no discernable effects, negative effects).

Although WWC has reviewed several interventions targeted to support students with disabilities (e.g., Dyslexia Training Program, Lovaas Model of Applied Behavior Analysis), the focus on most of the reviews has been broad educational programs rather than instructional strategies that can be incorporated into a teacher's daily routines and activities. Whereas the focus on broad educational programs is appropriate for general education, special education researchers have generally focused on determining whether discrete practices such as repeated readings (Chard et al., 2009), self regulated strategy development (Baker et al., 2009), and time delay (Browder et al., 2009) meet EBP standards.

Additionally, the WWC currently reviews only research studies with certain methodological designs (i.e., randomized control trials and quasi-experimental). Odom

(2005) explained that in order to understand effective practices in special education, other research methodologies should be included for evidence based reviews (e.g., single subject design). In the following section, I explain: (a) how establishing EBPs in special education differs from doing so in general education and (b) the importance for the special education researchers to develop their own standards and practices for establishing EBPs.

**Establishing EBPs in special education.** Over the past 30 years, special education research and policy has focused on improving the outcomes of students with disabilities and bridging the gap between research and practice (Greenwood & Abbott, 2001). Although instructional approaches and principles in general and special education overlap in many ways, special education is unique in its instructional focus. Although practices used by special educators to support students with disabilities may be effective for students without disabilities, students with disabilities *need* effective practices in order to succeed academically (Cook & Schirmer, 2006). Fuchs and Fuchs (1995) maintained that many practices validated by special education educators do not transfer easily into general education because special education practices focus on the individual student rather than using the same instructional strategies for an entire class. In addition, individualization is hardly observed in general education classes and, in reality, impractical for general education classrooms with 25-30 students (Fuchs & Fuchs). Special education is based on the premise of individualizing to meet the needs of each student with a disability and therefore special educators often teach students in small group settings or individually in order to meet students' individual needs.

Special education research has validated many effective practices unique to special education, and yet these practices are not frequently and consistently being used in the classrooms (Cook & Schirmer, 2003). In order to bridge the gap of research to practice in special education, Cook and Schirmer maintain that it is important for researchers to summarize, synthesize, expand, and update the research on effective practices for students with disabilities. In other words, to begin bridging the research to practice gap, special education practices need to undergo evidence based reviews.

In order to conduct evidence based reviews for discrete practices in special education, it will be necessary for researchers to expand reviews beyond broad educational programs as currently done by the WWC. The WWC tends to focus on broad practices targeting nondisabled learners and thus has reviewed relatively few discrete strategies targeting learners with disabilities. Indeed, the WWC has not examined the effectiveness of any practices for learners with low incidence disabilities. Because of the variability in special education participants (e.g., the Individuals with Disabilities Education Act specifies 13 categories of disabilities with varying characteristics) and their need for individualization, it will be important that EBP reviews in special education determine not only that the intervention is successful, but also *who* the intervention is successful for (Odom, 2008).

Due to the low incidence of many disabling conditions, special education researchers often conduct research studies using a small number of participants and employ research designs other group comparison designs. Specifically, in order to conduct evidence based reviews in special education, EBP standards should include the use of single subject research designs. This allows for research to be conducted on

interventions for students with low incidence disabilities (e.g., severe intellectual disabilities).

Although the WWC has reviewed thousands of studies in education, the Council for Exceptional Children (the largest and most influential organization devoted to children with exceptional needs in the world; CEC) has not formally adopted a process for determining EBPs for students with disabilities. In 2005, Gersten et al. and Horner et al. proposed the first set of standards for determining EBPs for group experimental and single subject research in special education. In 2009, several special education researchers piloted these standards (see Baker et al., 2009; Browder et al., 2009; Chard et al., 2009; Montague & Dietz, 2009; Lane et al., 2009;) and provided feedback on how the process for determining EBPs in special education could be improved. Most recently, Cook et al. (2013) have proposed revised a set of combined quality indicators and standards for group experimental, quasi-experimental, and single subject design that incorporates the feedback from the research reviews in 2009. The indicators and standards proposed by Cook et al. have yet to be field tested.

In the following section, I first describe research designs generally accepted for determining EBPs in special education. Next, I describe: (a) Gersten et al.'s (2005) proposed quality indicators for group comparison research (including randomized control trials and quasi-experimental design), (b) Horner et al.'s (2005) quality indicators for single subject design, and (c) Cook et al.'s (2013) quality indicators for group comparison research (including randomized control trials and quasi-experimental group designs) and single subject design. Then I will explain how each set of standards determines whether a practice can be considered an EBP.

***Research design.*** The first step in determining EBPs is to establish what types of research designs can be considered in an evidence based review. In special education, it is necessary for researchers to employ a variety of research designs in order to answer a variety of research questions (Odom et al., 2005). Group comparison, correlational, single subject, and qualitative designs are used by special education researchers. In the following, I provide a brief description of each type of research design.

*Group comparison research.* Group comparison research designs involve a researcher actively implementing an intervention (e.g., CWPT) on a group of individuals (i.e., treatment group) and not others (i.e., a control group) to examine whether differential change occurs in the dependent variable (e.g., math scores). Control groups are an essential component of group comparison designs. In order for a researcher to show that it was the independent variable (e.g., CWPT) that caused change in the dependent variable (e.g., math scores), it is necessary that the researcher use a control group (i.e., a group that does not receive the treatment) that is functionally equivalent to the group that receives the intervention.

Randomized controlled trials and quasi-experimental designs are two types of group experimental research. Randomized controlled trials are sometimes considered the "gold standard" of education research (WWC, 2011). When a researcher uses random assignment, he or she can control for extraneous variables (Odom, 2008) and prevent selection bias (Cook et al., 2008). In randomized control trials, the researcher randomly assigns participants to either the treatment or control group. Although this does not ensure that each group is exactly the same, by randomly assigning participants, the researcher can be sure that he or she is not adding bias to treatment assignment. The goal

of the researcher in group experimental design is to control for as many extraneous variables as possible in order to maximize the likelihood that the independent variable caused the change in the dependent variable.

In quasi-experimental group designs, the researcher uses intact groups that are already in place (Kennedy, 2005). The researcher will assign one or more groups (e.g., classrooms, schools) to the treatment group and one or more others to the control group. Although this design increases the potential of selection bias, the researcher can measure and balance differences in characteristics between groups that relate to the outcomes (Cook et al., 2008) by ensuring participants are matched across groups on important variables. In other words, the researcher should make both the control and treatment group as similar as possible on important variables. Alternatively, to equate groups in quasi-experimental group studies, the researcher may use a pre-assessment measure to determine differences between the groups and, if necessary, control for any differences statistically. Cook et al. (2013) consider regression discontinuity design a type of quasi-experimental research.

*Single subject research.* When using a single subject design, the researcher determines whether a functional relationship exists between the independent variable (i.e., intervention) and dependent variable (i.e., student outcome). Similar to group experimental design, if the researcher can demonstrate a functional relationship (i.e., functional control of the independent variable over the dependent variable), it can be reasonably inferred that the independent variable caused change in the dependent variable (Horner et al., 2005). In order for a researcher to show functional control he or she must replicate change multiple times within the study (i.e., the dependent variable must change

in the desired direction each time the independent variable is manipulated). Unlike group experimental design, single subject research does not require a control group. Individuals (or groups) serve as their own control. Using an ABAB design, for example, the researcher can show that the introduction and withdrawal of the independent variable causes consistent and meaningful changes in the dependent variable measure across multiple phases. Other single subject designs include (but are not limited to): (a) multi-element design, (b) multiple baseline design, and (c) combined designs (Kennedy, 2005).

All single subject designs share several commonalities. First, the researcher needs to operationally define participants, independent/dependent variables, and baseline and control conditions. Operational definitions are important because they allow other researchers to replicate the study. Replication is especially important in single subject design research because results of individual studies should not be generalized to the broader population (because of small *n*). Replication across multiple and diverse participants is therefore necessary for generalization.

*Correlational research.* Researchers using correlational research examine relationships between different variables (Thompson et al., 2005). For example, a researcher may examine the relationship between CWPT and math scores for middle school students with learning disabilities. Correlational research usually does not actively introduce an intervention (Cook et al., 2008) and therefore a researcher using correlational research cannot infer causality. For example, if a researcher comparing the performance change of two schools (School A was implementing CWPT and School B was not) found that School A outperformed School B the researcher could not infer that CWPT caused the higher performance. In order for a researcher to infer cause, an

intervention must be introduced and other variables, that could also cause increased school performance, be controlled for.

*Qualitative research.* Unlike group experimental and single subject design, the purpose of qualitative research is not to demonstrate causality. Qualitative research explores opinions, ideas, and beliefs (Bratlinger et al., 2005). Qualitative research is designed to answer "why" or "how" and gives insight into the lives of people with disabilities and those around them. Qualitative research designs do not fit a "one size fits all" model (Bratlinger et al.). In fact, qualitative researchers may adapt their methods during their data collection in order to collect the most valuable information. In qualitative research, the researcher does not typically implement an intervention, but observes in the natural setting (Cook et al., 2008). Qualitative researchers can collect data through: (a) observations, (b) interviews, (c) photographs, (d) existing documents (among others) (Bratlinger et al.). Whereas qualitative research does not directly contribute to whether a practice works, it does contribute to the knowledge base of EBPs. Specifically, qualitative research may provide (a) insight into why or how an intervention works; (b) information on the social validity of an intervention, and/or (c) insight into what components of an intervention may contribute either positively or negatively to its success (McDuffie & Scruggs, 2008).

*Designs for determining EBPs*. In order to determine whether an intervention is effective (e.g., does the intervention positively impact the academic outcomes of students?), some research designs are more appropriate than others. Specifically, research designs in which researchers can establish experimental control reliably determine the effectiveness of an intervention. Therefore, in special education, most researchers

generally agree that group comparison designs (including randomized control trials and quasi-experimental designs) and single subject research designs are appropriate for determining EBPs (see Cook et al., 2008; Gersten et al., 2005; Horner et al., 2005). Randomized control trials and quasi-experimental designs are commonly used for establishing what works in general education (see WWC, 2008). However, Odom (2008) suggested that single subject is an important design in special education because many individuals with disabilities are from low incidence populations. Because single subject research can involve a very small number of participants and also allows the researcher experimental control, it is has been accepted as a valid research design in determining EBPs for special education (Horner et al., 2005).

In 2005, Gersten et al. and Horner et al. proposed quality indicators and standards for identifying EBPs for group comparison (including randomized control trials and quasi-experimental designs) and single subject research, respectively. The 2005 quality indicators and standards were field tested by groups of special education researchers in a 2009 special issue of *Exceptional Children* (Cook, Tankersley, & Landrum; see also Stenhoff & Lignugaris/Kraft, 2007; Browder, Spooner, Ahlgrim-Delzell, Harris, & Wakeman, 2008; Test et al., 2009; Jitendra, Burgess, & Gajria, 2011). Cook et al. (2009) summarized the feedback from the field tests and concluded that quality indicators and standards should be more clearly defined and operationalized. Most recently, Cook et al. (2013) proposed a refined set of quality indicators and standards for determining EBPs in special education. In the next sections, I describe the 2005 and 2013 quality indicators and standards. Specifically, I review how each group of special education scholars proposed to (a) identify high quality studies (i.e., describe quality indicators of high

quality studies) and (b) determine whether an intervention can be considered an EBP (i.e., describe standards for determining EBPs).

**Determining high quality studies.** In order to determine EBPs in special education, a researcher must first target an intervention for review. The researcher then needs to locate all studies that use acceptable research designs and code all extant research using pre-determined quality indicators; this ensures that only high quality studies with research designs that allow experimenter control are included in the EBP review. In this section, I describe the 2005 quality indicators for group experimental research, including randomized control trials and quasi-experimental design (Gersten et al., 2005), and single subject design (Horner et al., 2005). Then, describe the 2013 quality indicators for group experimental (i.e., randomized control trials and quasi-experimental designs) and single subject design (Cook et al., 2013).

***Gersten et al.'s (2005) quality indicators for group experimental research.*** Gersten et al. (2005) proposed 10 essential quality indicators and 8 desirable indicators for group experimental research. Essential and desirable indicators are categorized into four topic areas: (a) describing participants, (b) description of intervention and comparison groups, (c) outcome measures, and (d) data analysis. Gersten et al. did not operationally define many of the quality indicators and suggested that the indicators needed to be refined in order to improve usability. In addition, Gersten et al. did not specify how the quality indicators should be rated (e.g., using a dichotomous scale or based on a rubric). In the following sections, I provide a description of the quality indicators for each topic area proposed by Gersten et al.

*Describing Participants.* Gersten et al. (2005) established three essential quality indicators for describing participants: (a) demonstration of disability, (b) comparability of groups, and (c) comparability of interventionists. Gersten et al. listed one desirable indicator for this area: attrition information. Information on participants and interventionists is important in a study because it provides information on generalization of a study (i.e., who may benefit from the practice, with what interventionists the practice may be successful).

In order to understand the population who benefits from receiving an intervention, it is important for researchers clearly describe participants of the study. Clearly identifying participants enables research consumers to interpret the findings of a study (e.g., for whom will benefit from the intervention). Researchers also need to clearly define the disability of participants by describing the criteria used in the disability determination; this allows research consumers to determine whether the participants actually experienced the disability. In order to meet the essential quality indicators participant description, Gersten et al. (2005) proposed that a researcher must address the following quality indicator:

- Was sufficient information provided to determine/confirm whether the participants demonstrated the disabilities or difficulties presented?

Gersten et al. (2005) determined that high quality studies should include detailed information on comparison/control groups and how participants are assigned to each group. In order to ensure that comparisons of pre-test/post-test differences between conditions are meaningful, it is important for researchers to ensure both treatment and control groups are similar on relevant variables. Gersten et al. suggested using random

assignment for comparison conditions whenever possible; however, quasi-experimental studies may be included in an evidence based review. In order to meet the essential quality indicators participant description, Gersten et al. proposed that a researcher must address the following quality indicator:

- Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?

To facilitate replication and ensure that the effects of the intervention were not due to effects associated with the interventionist, it is important for a researcher to include information about the person(s) providing the intervention and that interventionists are comparable across conditions. Furthermore, only qualified individuals should conduct an intervention. Gersten et al. also recommended that random assignment of interventionists is also preferred. In order to meet this essential quality indicator, Gersten et al. (2005) proposed that a researcher must address the following quality indicator:

- Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

Gersten et al. (2005) recommended one desirable indicator for describing participants. In order to meet this indicator, a researcher must provide information on attrition rates. In order to be confident that groups remain comparable from pre- to post-test, it is important for researchers to document and compare attrition rates for

intervention and control groups. In order to meet the desirable quality indicator for participant description, Gersten et al. proposed that a researcher address the following:

- Was data available on attrition rates among intervention samples? Was severe overall attrition documented? If so, is attrition comparable across samples? Is overall attrition less than 30%?

*Intervention and comparison conditions.* Gersten et al. (2005) proposed three quality indicators for describing and implementing the intervention and comparison conditions: (a) description of intervention, (b) implementation fidelity, and (c) comparison conditions.

In order for consumers to understand and potentially replicate the steps of an intervention it is important for researchers to clearly describe the intervention. In order to meet the essential quality indicator for intervention description, Gersten et al. (2005) proposed that a researcher address the following:

- Was the intervention clearly described and specified?

In order to understand the relationship between the intervention and outcome measures, it is important that the researcher determine that the intervention was implemented as intended (Gersten et al., 2005). To do this, a researcher must measure implementation fidelity and describe how it was measured (Gersten et al., 2005). Gersten et al. also recommends the research team include a measure of inter-observer reliability when measuring implementation fidelity. In order to meet the essential quality indicator for participant description, Gersten et al. (2005) proposed that a researcher must address the following:

- Was the fidelity of implementation described and assessed?

Additionally, Gersten et al. (2005) suggested that it is desirable to go beyond the researcher solely documenting implementation fidelity. In order to provide a deeper understanding about the implementation issues of an intervention and to gain insight about intervention components, a researcher should document the quality of implementation (Gersten et al., 2005, p. 157). Gersten et al. proposed two desirable quality indicators for capturing the quality of the intervention:

- Did the research team assess not only surface features of fidelity of implementation (e.g., # number of minutes allocated to CWPT or teacher following procedures specified), but also examine the quality of implementation?
- Did the research report include actual audio or videotape excerpts that capture the nature of the intervention?

The third essential quality indicator refers to the importance of describing the comparison condition (i.e., describe nature of services in the comparison conditions). In order "to understand what an obtained effect means, one must understand what happened in the comparison classrooms" (Gersten et al., 2005, p. 158). In order to meet the essential quality indicator for the comparison condition, Gersten et al. proposed that a researcher address the following:

- Was the nature of services provided in comparison conditions described?

Additionally, Gersten et al. (2005) suggested that it is desirable for researchers to document the nature of instruction provided in the comparison condition. In order to increase understanding of comparison conditions, a researcher should document the specific instruction that occurred in the comparison condition. In order to meet this

desirable quality indicator, Gersten et al. (2005) proposed that a researcher address the following:

- Was any documentation of the nature of instruction or series provided in comparison conditions?

*Outcome measures*. In order to ensure validity of outcomes, it is important for a researcher to measure outcomes appropriately. Gersten et al. (2005) outlined two essential indicators for this category: (a) multiple measures and (b) appropriate outcomes.

In order to increase confidence that the independent variable affected the dependent variable, it is important for a researcher to use multiple outcome measures. Gersten et al. (2005) explained that no one measure can assess all aspects of an outcome and it is valuable to have "multiple tools to measure each facet of performance" (p.158). In addition, it is important that a researcher considers using measures of generalized performance rather than measures too closely aligned with the intervention in order to prevent "teaching to the test." In order to meet this essential quality indicator for outcome measures, Gersten et al. proposed that a researcher address the following:

- Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance?

In addition to using multiple measures, Gersten et al. (2005) maintained that outcomes of the intervention must be measured at appropriate times in order to determine changes in the dependent variable. In some cases, outcomes measures should be measured immediately, whereas at other times it is important to wait to collect outcome

data. In order to meet this essential quality indicator for outcome measures, Gersten et al. (2005) proposed that a researcher must the following:

- Were outcomes for capturing the intervention's effect measured at the appropriate time?

Additionally, Gersten et al. (2005) recommended that it is desirable for researchers to collect data beyond an immediate post-test. In order to provide information on an intervention's long term effects, it is important researchers measure outcomes beyond a single post-test. In order to meet this desirable quality indicator for outcome measures, Gersten et al. proposed that a researcher must address the following:

- Were the outcomes for capturing the intervention's effect measured beyond an immediate post-test?

It is also desirable that a researcher consider evidence of (a) internal consistency reliability and (b) data collection activities (Gersten et al., 2005). In order to understand how well items on a test (i.e., outcome measure) fit together, it is important for researchers to provide information on internal consistency. Additionally, in order for researchers to prevent biased data collection, it is important to keep data collectors unaware of study conditions. In order to meet this desirable quality indicator for outcome measures, Gersten et al. (2005) proposed that a researcher address the following:

- Did the study provide not only internal consistency reliability but also test-retest reliability and inter-rater reliability (when appropriate) for outcome measures? Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?

Documenting construct validity is also desirable for a research study. Gersten et al. (2005) suggested that in order for a study to be ranked highly acceptable, a researcher should include data on predictive validity of measures and information on construct validity. In order to meet this desirable quality indicator for outcome measures, Gersten et al. proposed that a researcher address the following:

- Was evidence of the criterion-related validity and construct validity of the measures provided?

*Data analysis*. Gersten et al. (2005) proposed two essential quality indicators for data collection: (a) appropriate data analysis techniques and (b) effect size calculations. It is important for researchers to use appropriate data analysis techniques to meaningfully address the research questions. In order to meet this essential quality indicator for data analysis, Gersten et al. proposed that a researcher must address the following:

- Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?

Gersten et al. (2005) suggested that it is important for a researcher to provide effect size calculations in addition to providing inferential statistics. Effect sizes allow the reader to understand the amount of impact an intervention had on the treatment group. In order to meet this essential quality indicator for data analysis, Gersten et al. proposed that a researcher address the following:

- Did the research report include not only inferential statistics but also effect size calculations?

Lastly, Gersten et al. (2005) suggested that it is desirable for researchers to present results in a clear and coherent fashion. In order to meet this desirable quality indicator for data analysis, Gersten et al. proposed that a researcher must address the following:

- Were results presented in a clear coherent fashion?

Gersten et al. (2005) proposed that reviewed studies should be labeled as high quality or acceptable according to the number of essential and desirable indicators met. In order for a group experimental study to be considered high quality, the study must meet: (a) all but one of the essential quality indicators and (b) at least four of the desirable quality indicators. To be considered acceptable a study must meet (a) all but one of the essential quality indicators and (b) at least one of the desirable quality indicators (see Table 3).

***Horner et al.'s (2005) quality indicators for single subject research.*** Horner et al. (2005) proposed 21 essential indicators for single subject research. The 21 indicators can be categorized into seven different topic areas: (a) participants and setting, (b) dependent variable, (c) independent variable, (d) baseline/comparison condition, (e) experimental control/internal validity, (f) external validity, and (g) social validity. Unlike Gersten et al.'s (2005) quality indicators for group design, all 21 quality indicators described by Horner et al. are considered essential when implementing a high quality single subject study. Horner et al. did not specify how presence of quality indicators should be assessed (e.g., dichotomous variable or rubric).

*Describing participants and setting.* Horner et al. (2005) described 3 quality indicators for describing participants and setting: (a) participant description, (b)

participant selection, and (c) description of the physical setting. In order to determine confidently with whom and where the intervention is effective, it is important for researchers to describe well the participants and setting of the study. In order to meet the three indicators for participants and setting, Horner et al. proposed that a researcher address the following:

- Are participants described with sufficient detail to allow other to select individuals with similar characteristics (e.g., age, gender, disability, diagnosis)?
- Was the process for selecting participants described with replicable precision?
- Were critical features of the physical setting described with sufficient precision to allow replication?

*Dependent variable.* "Single subject research employs one or more dependent variables that are defined and measured" (Horner et al., 2005, p. 167). It is important for researchers to describe the dependent variable within a study in order for the researcher to consistently assess and replicate the assessment process across phases. Additionally, the dependent variable must be measured repeatedly across phases in order for the researcher to compare the performance of the participant during baseline and intervention phases. Horner et al. described five quality indicators emphasizing the importance of dependent variable(s):

- Were dependent variables described with operational precision?
- Was each dependent variable measured with a procedure that generates a quantifiable index?

- What the measurement of the dependent variable valid and described with replicable precision?
- Were dependent variables measured repeatedly over time?
- Was data collected on the reliability or inter-observer agreement associated with each dependent variable, and IOA levels meet minimal standards (e.g., IOA = 80%; Kappa = 60%)

*Independent variable.* In single subject research the independent variable is typically an intervention, practice, or behavioral mechanism (Horner et al., 2005). In order to allow a research consumer to interpret the results of the study and for replication purposes, the independent variable must be described in detail. The researcher must also determine how to implement the independent variable in order to document experimental control. Specifically, the researcher must actively manipulate the independent variable and provide evidence (through visual data) that the independent variable affected the dependent variable in the desired direction. Horner et al. established three quality indicators for the independent variable in single subject design: (a) description of the independent variable, (b) manipulation of the independent variable, and (c) fidelity of implementation. In order to meet the three indicators, Horner et al. proposed that a researcher address the following:

- Was the independent variable described with replicable precision?
- Was the independent variable systematically manipulated and under the control of the experimenter?
- Was overt measurement of the fidelity of implementation for the independent variable highly desirable?

*Baseline.* The baseline condition in single subject research is comparable to comparison/control group in group experimental research; single subject researchers compare effects of the independent variable to performance during baseline condition (Horner et al., 2005). The baseline refers to the condition when the independent variable has not yet been introduced. Horner et al. (2005) indicated two essential quality indicators for baseline condition: (a) description of baseline and (b) collection of baseline data. Baseline conditions are comparable to control groups in group design research. Baseline conditions need to be clearly established in order to compare performance during treatment conditions. The baseline condition should be described in enough detail to allow replication by other researchers (Horner et al.). In order to meet the two indicators, Horner et al. proposed that a researcher address the following:

- Did the baseline phase provide repeated measurement of a dependent variable and establish a pattern of responding that can be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur?
- Were baseline conditions described with replicable precision?

*Experimental control/ internal validity.* Unlike group design, single subject researchers do not use random assignment or comparison groups to show experimental control. Instead, when repeated manipulation of the independent variable corresponds reliably with changes the dependent variable (in the desired direction), functional or experimental control is established. Experimental control, in single subject research, can be established by (a) introducing and withdrawing the independent variable, (b) staggering the introduction of the independent variable at different times, or (c)

repeatedly manipulating the independent variable across different observational periods (Horner et al., p.168). Researchers employing multiple baseline designs must document experimental control by providing a single subject graph that allows evaluation of changes in level, trend, and variability in performance across all phases. Horner et al. (2005) suggested three indicators related to experimental control: (a) demonstration of experimental effect, (b) controlling for threats to internal validity, and (c) documentation of experimental control. In order to meet the three indicators, Horner et al. proposed that a researcher address the following:

- Did the design provide at least three demonstrations of experimental effect at three different points in time?
- Did the design control for common threats to internal validity (e.g., permits elimination of rival hypotheses)?
- Did results document a pattern that demonstrates experimental control?

*External validity.* External validity refers to the extent to which the results of the study can be generalized to other participants in other locations. Unlike group experimental and quasi-experimental research, single subject research studies involve a small number of participants. Hence, it is expected that single subject researchers replicate effects with other participants, settings, or both to increase external validity (Horner et al., 2005). In order to meet the indicator for external validity, Horner et al. (2005) proposed that a researcher address the following:

- Were experimental effects replicated across participants, settings, or materials to establish external validity?

*Social validity.* In single subject research, it is important for researchers to establish a study's social validity. Social validity refers to the impact the intervention will have on the participant(s) of the study. To ensure that researchers choose to implement meaningful interventions, they should document social validity. Horner et al. (2005) suggested four indicators for social validity: (a) importance of dependent variable, (b) magnitude of change in dependent variable, (c) cost effectiveness of intervention, and (d) realistic implementation. In order to meet the four indicators for social validity, Horner et al. proposed that a researcher address the following:

- Was the dependent variable socially important?
- Was the magnitude of change in the dependent variable resulting from the intervention socially important?
- Was the implementation of the independent variable practical and cost effective?
- Was social validity enhanced by the implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts?

In order for a study to be included in an evidence based review, a study should meet all 21 quality indicators proposed by Horner et al. (2005). Therefore, if a study fails to meet one indicator proposed by Horner et al., it should not be considered methodologically sound and cannot be included in determining whether the practice is evidence based.

***2013 quality indicators.*** Using the previous work done by Gersten et al. (2005), Horner et al. (2005), and the feedback from the 2009 special issue of *Exceptional*

*Children*, the CEC Workgroup proposed a combined set quality indicators for both group experimental (true experimental and quasi-experimental design) and single subject design. The CEC Workgroup intended that for each of the quality indicators, a study should be rated on a dichotomous scale: met or not met. Specifically, the quality indicator can be met if the study under review reasonably addressed the quality indicator; studies do not need to completely or absolutely meet the quality indicator (Cook et al., 2013). In order to be included in an EBP review, studies must be of strong methodological quality. Strong methodological studies must meet all of the quality indicators related to the research design used.

Cook et al. (2013) proposed 31 essential quality indicators for group comparison research (i.e., randomized control trials and quasi-experimental designs) and single subject research. These 31indicators can be categorized into nine different topic areas: (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) implementation fidelity, (f) internal validity, (g) outcome measures/dependent variables, and (h) data analysis. Nineteen of the quality indicators proposed by Cook et al. (2013) apply to both group comparison research and single subject research designs (e.g., relevant demographics provided for participants, inclusion/exclusion criteria for study eligibility are provided). Eight quality indicators apply only to group comparison research design (e.g., description of how context/settings were selected) and four quality indicators apply only to single subject design (e.g., single subject design controls for common threats to internal validity).

Cook et al. (2013) invited 25 expert special education researchers to participate in a Delphi study in order to: (a) gain input for refining the 2013 quality indicators and

criteria and (b) find consensus regarding the acceptability of the standards. In order to participate in the Delphi study, participants (a) were nominated by a CEC Workgroup member as an expert special education researcher in group comparison research, single subject research, or both; (b) had published at least two group comparison or single subject research studies in prominent special education research journals since 2000; and (c) were unanimously approved by all Workgroup members. Out of the 25 researchers invited, 24 researchers agreed to participate. One expert who originally agreed to participate did not due to illness; therefore there were a total of 23 participants. At the start of the study, participants had an average of 14.1 years experience as special education researchers (*SD* = 6.7, range 7-30), published an average of 5.9 group comparison studies (*SD* = 5.8, range 0-25), and published an average of 8.2 single subject studies (*SD* = 9.8, range 0-40).

Delphi Study participants were asked to rate on a 1 to 4 scale (where 1 is strongly disagree, 4 is strongly agree): (a) each area of quality indicators and each evidence-based classification; (b) completeness of quality indicators and evidence-based classifications; (c) descriptions of research designs considered; (d) criteria for sound studies; and (e) classification of study effects for group comparison and single-subject studies. Additionally, participants were asked to give a reason for their rating and provide suggestions for improvement in each area. The CEC Workgroup considered feedback from Delphi participants and incorporated Delphi participant input that the majority of the Workgroup agreed as beneficial until a minimum of 80% of Delphi participants agreed (rating of 3) or strongly agreed (rating of 4) for all areas rated.

In Round 1 of the Delphi study, participants' mean rating was 3.5 (*SD* = 0.7) with a median and modal rating of 4. The overall percentage of agree/strongly agree ratings was 91%. The lowest rated areas included:

- Description of how research designs and methodological quality of studies were used in determining the evidence-based classification of practices, which included noting that moderate quality studies were weighted as half of a methodologically strong study (*M*=3.0, *SD*=0.9, % agree=71%)
- Criteria for determining positive, neutral, and negative effects for single-subject studies (*M*=3.1, *SD*=0.9, % agree=65%)
- Description of the quality indicators, which included noting that (a) items related to social validity, implementation fidelity, and reporting effect size (or data from which effect sizes can be calculated) are not required for moderate quality studies and (b) a quality indicator is met "when the study under review reasonably addresses the spirit of the quality indicator" (*M*=3.3, *SD*=0.9, % agree=83%)
- Quality indicators for intervention agents, which required that intervention agents be equivalent across conditions (*M*=3.3, *SD*=0.9, % agree=83%)

After incorporating changes as a result of feedback from Delphi participants as well as recommendations from Workgroup members, Cook et al. (2013) continued another round of Delphi study to seek feedback and approval of the revised quality indicators and standards. In Round 2, the mean rating from Delphi participants was 3.8 (*SD* =0.5) across all areas, which indicated an improved mean rating and less variability in comparison to Round 1. As in Round 1, the median and modal rating was 4. The

overall percentage of agree/strongly agree ratings was 98%. Twenty-two of 23 participants agreed or strongly agreed with 7 of the areas (96%); all participants agreed or strongly agreed with the remaining 11 areas. Thus, no area fell below the threshold of 80% agreement. Cook et al. incorporated the additional feedback from Round 2 of the Delphi study in order to finalize the 2013 quality indicators and standards for determining EBPs in special education.

In following sections, I provide a description of the 2013 quality indicators for each topic area developed (and refined from Delphi study feedback) by the CEC Workgroup for group comparison research and single subject research. Quality indicators apply to both research designs unless otherwise noted.

*Context and setting.* In order for a study to be considered an EBP, the researcher must provide sufficient information regarding the context or setting in which the intervention occurred. In order to meet the quality indicator for context and setting, Cook et al. (2013) proposed that a researcher must address the following:

- Characteristics of the critical features of the context(s) or setting(s) relevant to the study (e.g., type of program[s]/classroom[s], type of school [e.g., public, private, charter, preschool], curriculum used, geographical location[s], community setting[s], socio-economic status, physical layout[s]) are described.

*Participants.* Cook et al. (2013) established two quality indicators for describing participants: (a) description of demographics and (b) disability or risk status of participants. In order to meet the quality indicators, Cook et al. proposed that a researcher must address the following:

- Participant demographics relevant to the study (e.g., gender, age/grade, race/ethnicity, socio-economic status, language status) are described.
- Disability or risk status of participants (e.g., specific learning disability, autism spectrum disorder, behavior problem, at-risk for reading failure) and method for determining status (e.g., identified by school using state IDEA criteria, teacher nomination, standardized intelligence scale, curriculum-based measurement probes, rating scale) are described.

*Intervention agents.* In order to meet the two quality indicators for describing intervention agents, it is important for researchers to provide sufficient information characterizing the interventionists and their training or qualifications related to the intervention (Cook et al., 2013). In order to meet the two indicators, Cook et al. (2013) proposed that a researcher must address the following:

- Role (e.g., teacher, researcher, paraprofessional, parent, volunteer, peer tutor, sibling, technological device/computer) and background variables of intervention agent(s) relevant to the study (e.g., gender, race/ethnicity, educational background/licensure, professional experience, experience with intervention) are described.
- If specific training (e.g., amount of training, training to a criterion) or qualifications (e.g., professional credential) are required to implement the intervention, they are described and achieved by interventionist(s).

*Description of practice.* Cook et al. (2013) maintained that sufficient information must be provided regarding the critical features of the practice so the practice is clearly

understood and can be reasonably replicated. In order to meet the two quality indicators, Cook et al. (2013) proposed that a researcher must address the following:

- Detailed intervention procedures (e.g., intervention components, instructional behaviors, critical or active elements, manualized or scripted procedures, dosage) and intervention agents' actions (e.g., prompts, verbalizations, physical behaviors, proximity) are described, or one or more accessible sources are cited that provide this information.
- When relevant, materials (e.g., manipulatives, worksheets, timers, cues, toys) are described, or one or more accessible sources are cited that provide this information.

*Implementation fidelity*. Cook et al. (2013) described three quality indicators of acceptable implementation fidelity. In order to meet the quality indicators, Cook et al. (2013) proposed that a researcher must address the following:

- Implementation fidelity related to adherence is assessed using direct, reliable measures (e.g., observations using a checklist of critical elements of the practice) and reported.
- Implementation fidelity related to dosage or exposure is assessed using direct, reliable measures (e.g., observations or self-report of the duration, frequency, and/or curriculum coverage of implementation) and reported.
- Implementation fidelity is (a) assessed regularly throughout implementation of the intervention (e.g., beginning, middle, end of the intervention period) as appropriate for the study being conducted; (b) assessed for each interventionist, each setting, and each treatment group

(or participant in single subject-research in which individuals are the unit of analysis) as relevant, and (c) reported.

*Internal validity*. Cook et al. (2013) described four areas of importance for internal validity: (a) experimenter control, (b) nature of services in comparison conditions, (c) evidence independent variable changing dependent variable, and (d) attrition. In order to ensure the independent variable is under the control of the experimenter, Cook et al. proposed a researcher should address the following quality indicator:

- The researcher(s) controls and systematically manipulates the independent variable.

Cook et al. (2013) recommended that researchers must provide detailed information regarding the control/ comparison group in order to prevent threats to internal validity. Cook et al. proposed that a researcher address the following:

- The curriculum, instruction, and interventions used in control/comparison condition(s) (in group comparison studies) or baseline/comparison phases (in single-subject studies) are described (e.g., definition, duration, length, frequency, learner:instructor ratio).
- Access to the treatment intervention by control/comparison group(s) (in group comparison studies) or during baseline/comparison phases (in single-subject studies) is not provided or is extremely limited.

In both group experimental research and single subject research, it is important the research design provides sufficient evidence that the independent variable cause

change in the dependent variable. In order to meet the quality indicator(s), Cook et al. (2013) proposed that a researcher must address the following:

- For group comparison research:
    - Assignment to groups is clearly and adequately described.
    - Participants (or classrooms, schools, or other unit of analysis) are assigned to groups in one of the following ways: (a) randomly; (b) nonrandomly, but the comparison group(s) is matched very closely to the intervention group (e.g., matched on prior test scores, demographics); (c) non-randomly, but techniques are used to measure and, if meaningful differences (e.g., statistically significant difference, difference of > 0.05 pooled SDs) are identified, statistically control for any differences between groups on relevant pre-test score and/or demographic characteristics (e.g., statistically adjust for confounding variable through techniques such as ANCOVA or propensity score analysis); or (d) non-randomly on the basis of a reasonable cutoff point when regression discontinuity design is used.
- For single subject research:
    - The design provides at least three demonstrations of experimental effects at three different points in time.
    - For single-subject research designs that use a baseline phase, baseline phase includes at least three data points (except when fewer are justified by study authors due to reasons such as [a]

measuring severe and/or dangerous problem behaviors and [b] zero baseline behaviors with no likelihood of improvement without intervention) and establishes a pattern that predicts undesirable future performance (e.g., increasing trend in problem behavior, consistently infrequent exhibition of appropriate behavior, highly variable behavior).

- The design controls for common threats to internal validity (e.g., ambiguous temporal precedence, history, maturation, diffusion) such that plausible, alternative explanations for findings can be reasonably ruled out. Commonly accepted designs such as reversal (ABAB), multiple baseline, changing criterion, and alternating treatment address this quality indicator when properly designed and executed, although other approaches can be accepted if the researcher(s) justifies how they rule out alternative explanation for findings/control for common threats to internal validity.

When using group comparison research, it is important for researchers to ensure attrition was not a significant threat to internal validity. In order to meet the two quality indicators, Cook et al. (2013) proposed that a researcher must address the following when using group experimental design (i.e., these quality indicators do not apply to single subject designs):

- Overall attrition is low across groups (e.g., < 20% in a one-year study).

- Differential attrition (between groups) is low (e.g., within 20% of each other) or is controlled for by adjusting for non-completers (e.g., conducting intent-to-treat analysis).

*Outcome measures/dependent variables*. Cook et al. (2013) outlined seven indicators for outcome measures in two different areas: (a) applying appropriate measures and (b) demonstrating adequate psychometrics. In order to meet the four quality indicators related to the application of outcome measures, Cook et al. (2013) proposed that a researcher address the following:

- The outcome(s) is socially important (e.g., it constitutes or is theoretically or empirically linked to improved quality of life, an important developmental/ learning outcome, or both).
- Measurement of the dependent variable(s) is clearly defined and described.
- The effects of the intervention on all measures of the outcome(s) targeted by the review are reported (p levels and effect sizes [or data from which effect sizes can be calculated] for group comparison studies; graphed data for single-subject studies), not just those for which a positive effect is found.
- Frequency and timing of outcome measures are appropriate. For group comparison studies, outcomes must be measured at both pre- and post-test at a minimum. For single-subject studies, a minimum of 3 data points per phase must be measured (except when fewer are justified by study authors due to reasons such as [a] measuring severe and/or dangerous problem

behaviors and [b] zero baseline behaviors with no likelihood of improvement without intervention).

Additionally, Cook et al. (2013) suggested that outcomes measures must demonstrate adequate psychometrics. In order to meet the quality indicators, Cook et al. (2013) proposed that a researcher address the following:

- Adequate evidence of internal reliability, inter-observer reliability, test-retest reliability, and/or parallel form reliability, as relevant, is described (e.g., score reliability coefficient > .80, IOA > 80%, or Kappa > 60%).
- For group experimental designs:
    - Adequate evidence of concurrent, content, construct, or predictive validity is described (e.g., a specific validity coefficient is reported).
    - Evidence of reliability and validity (with the exception of inter-observer reliability, which must be evaluated using data within the study) are empirically evaluated based on (a) data generated within the study (i.e., researchers use their own data) or (b) data from another study. If evidence is imported from another study, the sample and scores are similar enough to make generalization to the current study sensible.

*Data analysis.* In order to meet the four quality indicators proposed by Cook et al. (2013) the researcher must: (a) appropriately conduct data analysis and, for group experimental designs, (b) report effect size calculations.

In both group comparison research and single subject research, it is important for researchers to conduct appropriate data analysis procedures. In order to meet the quality indicators, Cook et al. (2013) proposed that a researcher must address the following:

- Data analysis techniques are appropriately linked to the unit of analysis in the study. For example, if classrooms are randomly assigned to conditions in a group comparison study, then classroom (not individual) should be the unit of analysis (with the exception of multilevel analyses such as HLM, in which multiple units of analysis exist). Similarly, if the research question for a single-subject study is stated in terms of the effect of an intervention on a classroom, then classroom-level data should be analyzed.
- For single subject research:
    - A single-subject graph clearly representing outcome data across all study phases is provided for each unit of analysis (e.g., individual, classroom or other group of individuals) so that reviewers can determine the effects of the practice (see Classifying Effects of Studies section of this document). Regardless of whether study authors include their own visual or other analyses of data, graphs depicting all relevant dependent variables targeted by the review should be clear enough for reviewers to draw basic conclusions about experimental control using traditional visual analysis techniques (i.e., analysis of mean, level, trend, overlap, and consistency of data patterns across phases).
- For group comparison designs:

- Data analysis techniques are appropriate for comparing change in performance of two or more groups (e.g., t-tests, (M)ANOVAs, (M)ANCOVAs, hierarchical linear modeling, structural equation modeling). If atypical procedures are used, a rationale justifying the data analysis techniques is provided.

Additionally, Cook et al. (2013) suggested that reporting effect size is important in determining whether a study can be considered high quality in group comparison research. In order to meet the quality indicators, Cook et al. proposed that a researcher using group experimental research design, address the following:

- One or more appropriate effect size statistics (e.g., Cohen's d, Hedge's G, Glass's Δ, eta-squared) is reported for each primary outcome, even if the outcome is not statistically significant; or data are provided from which appropriate effect sizes can be calculated.

**EBP standards.** After determining whether studies meet the proposed quality indicators, the next step in determining whether an intervention is an EBP is to determine if the high quality studies meet the standards of an EBP. In the next sections, I will describe the standards proposed by (a) Gersten et al. (2005) for group experimental research studies; (b) Horner et al. (2005) for single subject studies; and (c) Cook et al. (2013) for both group comparison and single subject research studies.

*Gersten et al.'s (2005) standards for evidence-based practices.* Gersten et al. (2005) proposed two categories for effective practices as determined by group experimental research: EBPs and promising practices. In order for an intervention to be considered an EBP, Gersten et al. suggested that there must be (a) two high quality

studies or four acceptable studies that support an intervention and (b) the weighted effect size must be significantly greater than zero. In order for an intervention to be determined a promising practice, there must be: (a) at least four acceptable quality studies or two high quality studies and (b) "there is a 20% confidence interval for the weighted effect size that is greater than zero" (Gersten et al., 2005, p. 162). Gersten et al. did not suggest labels for interventions not meeting standards for EBP or promising practice.

*Horner et al.'s (2005) standards for evidence-based practices.* In order for an intervention to be considered an EBP on the basis of single-subject research, Horner et al. (2005) proposed five standards. Four of the standards (i.e., intervention must be well defined in order to be replicated; the researcher must describe conditions for intervention, qualifications of who can implement the intervention, and who the intervention is effective for; intervention was implemented with documented fidelity; functional relationship), refer to quality indicators that each study included in the evidence based review must meet. Therefore, if studies were reviewed appropriately, the intervention will necessarily meet the first four standards by meeting the five criteria.

In order to meet the fifth standard proposed by Horner et al. (2005), and be determined an EBP, the high quality single subject studies must meet the following criteria:

- A minimum of five single subject studies that meet all 21 quality indicators and are published in peer reviewed journals
- The studies have been conducted by at least three different researchers in three different geographical locations
- The participants in the five studies must include at least 20 participants

***Cook et al. (2013).*** Cook et al.'s (2013) combined standards for group experimental and single subject research studies categorized an intervention's evidence into the following categories: (a) EBPs, (b) potential EBP, (c) insufficient evidence, (d) mixed evidence, and (e) negative effects.

In order for an intervention to be considered an EBP, intervention studies must meet the following criteria:

- At least two methodologically sound experimental group studies with positive effects and at least 60 participants across studies or
- At least 4 methodologically sound quasi-experimental (i.e. not randomly assigned) group studies with positive effects and at least 120 participants OR
- At least 5 methodologically sound single subject studies with positive effects and at least 20 total participants across studies OR
- At least 50% of the criteria for two or more of the study designs described above. For example, CWPT may be considered an EBP if there is one methodologically sound group experiment (with random assignment, positive effects, and at least 30 participants) and three high quality single subject research studies (with positive effects and at least 10 total participants).

In addition to the above criteria, Cook et al. also suggested that, in order for an intervention to be considered an EBP, there must: (a) be no high quality studies with negative effects and (b) a ratio of at least 3:1 of high quality studies with positive effects to high quality studies with neutral effects. For this criterion, Cook et al. considered group experimental, quasi-experimental, and single subject studies collectively.

For a potential EBP, Cook et al. (2013) proposed the following criteria:

- At least one methodologically sound experimental study with positive effects

OR

- At least 2 methodologically sound quasi-experimental (i.e. not randomly assigned) studies with positive effects OR
- At least 2 to 4 methodologically sound single subject studies with positive effects OR
- At least 50% of the criteria for two or more of the study designs described above. For example, CWPT may be considered a potential EBP if there is one high quality single subject research study (with positive effects) and one methodologically sound quasi-experimental group comparison study (with positive effects).

In addition to the above criteria, Cook et al. (2013) also suggested that, in order for an intervention to be considered a potential EBP, there must be: (a) no methodologically sound studies with negative effects and (b) a ratio of at least 2:1 of methodologically sound studies with positive effects to methodologically sound studies with neutral effects. For the criterion of potential EBP, Cook et al. do not require a minimum number of participants across studies.

For interventions that do not meet standards of an EBP or potential EBP, Cook et al. (2013) proposed criteria for interventions with: (a) insufficient evidence, (b) mixed evidence, or (c) negative effects. An intervention is labeled as having insufficient evidence if:

- An insufficient number of methodologically sound studies exist to meet any of the other evidence based categories

Cook et al. (2013) proposed that a research base is labeled as having mixed evidence if:

- Methodologically sound studies meet criteria for an EBP or a potential EBP AND
- One or more methodologically sound studies has negative effects, but the methodologically sound studies with negative effects do not outnumber methodologically sound studies with positive effects OR
- A ratio of methodologically sound studies with positive effects to methodologically sound studies with neutral effects is less than 2:1

A research base is labeled as having negative evidence if:

- One or more methodologically sound studies with negative effects AND
- The number of methodologically sound studies with negative effects outnumber the number of high quality studies with positive effects

Since the publication of Gersten et al. (2005) and Horner et al. (2005) several research teams (e.g., Browder et al., 2006, Baker et al., 2009, Chard et al., 2009) have used the quality indicators to determine the EBP status of interventions frequently used in special education (e.g., repeated reading, cognitive strategy instruction, self-regulated strategy development). However, EBP reviews have not (a) been conducted on all interventions that are recommended for students with disabilities and (b) been conducted using the 2013 quality indicators. In particular, CWPT is one intervention that is regarded in both research and practice as an effective intervention for students with disabilities, yet researchers previously have not formally reviewed its evidence based

status. In the next section, I define CWPT and the research that has been conducted on CWPT for students with disabilities.

**Classwide Peer Tutoring (CWPT)**

CWPT is an instructional strategy that is based on reciprocal peer tutoring and group reinforcement. The entire classroom of students, when participating in CWPT, is actively engaged in learning and practicing basic skills in a game based format (Terry, n.d.). CWPT was developed in the local schools of Kansas City, Kansas (Greenwood, 1997) in 1980 (Greenwood, Delquadri, & Carta, 1997). CWPT was developed to increase literacy of children who are poor, have mild disabilities, and who are culturally diverse (Greenwood, 1997). CWPT has been researched for over 31 years and it has been reported that the peer tutoring process increases on-task performance and academic achievement (Terry, n.d.) in reading (e.g., Kamps. Barbetta, Leonard, & Delquadri, 1994), spelling (e.g., Maheady & Harper, 1987), math (e.g. Maheady, Sacca, & Harper, 2001), science (e.g. Mastropieri et al., 2006), and social studies (e.g. Maheady, Harper, & Sacca, 1988a). CWPT has also been implemented in elementary, middle, and high school for students in general education, special education, and English Language Learners (ELL) (Terry, n.d). In the following sections, I (a) describe CWPT procedures and (b) review the CWPT research for students with disabilities.

**CWPT procedures.** CWPT is an instructional strategy to increase student engagement during instruction. In the following, I describe the essential elements associated CWPT: (a) assigning students to pairs and teams, (b) moving to tutoring positions, (c) passing out materials, (d) roles of the tutor/tutee and teacher during CWPT session, and (e) keeping points. These procedures are defined by the creators of CWPT

and outlined in their manual (Greenwood et al., 1997); some teachers may choose adjust CWPT procedures to meet the needs of their students and content area (e.g., choosing to implement CWPT sessions for more than 20 minutes per student). Teachers can also use the CWPT manual to support them with the CWPT procedures. In addition to the basic steps of CWPT, the manual provides additional information regarding:

- How to explain to students what peer tutoring is and how to be a good sport.
- How to adapt CWPT for individual subjects (e.g., spelling, math, reading fluency).
- How to troubleshoot certain issues (e.g., students off-task, odd number of students, students not following procedures).

To begin using CWPT, teachers arrange students into pairs (Greenwood et al., 1997). Greenwood (1997) recommended that pairs be switched on a weekly basis in order to keep things interesting and ensure that students are able to work with many classmates over time. Teachers may choose to pair students randomly or use skill pairing. Skill pairing refers to choosing students that have equal abilities to work together or choosing a student with higher skills to work with a student with lower skills (Greenwood et al., 1997). Teacher will designate one student in the pair as the "mover" and the other student as the "stayer." These role assignments inform students which partner will move to the other during CWPT. Assigning movers and stayers assists the teacher in maintaining a smooth transition into the CWPT session (i.e., students know exactly where to go and who to sit next to). Teachers should have a poster specifying pairs, movers, and stayers posted prominently in the classroom.

After pairs are chosen, the teacher divides the class into two teams; tutor pairs must be on the same team (Greenwood et al., 1997). Teams can be chosen randomly and the teacher can choose to allow students to give their team a name (Greenwood et al., 1997). Teams should also change weekly. See Appendix A for an example of a team and partner chart.

To begin a CWPT session, the teacher has students move to tutoring positions ("mover moves to their partner, the stayer"). Next, the teacher passes out the tutoring materials (i.e., what teachers want tutors/tutees to practice), tutoring point sheets (see Appendix B), and help signs. The teacher sets a certain amount of time (approximately 10 minutes) for the first part of the session. During the first 10 minutes, one student will be the tutor. The tutor will ask questions (e.g., spelling words, vocabulary) to the tutee and record 2 points for each correct answer on the tutoring point sheet. If a tutee gets an answer incorrect, the tutor corrects the tutee. The tutee can earn one point if the tutee writes and says the correct answer three times (Maheady, Harper, & Sacca, 1988a). After 10 minutes, the teacher resets the time for 10 more minutes and the pair will switch roles. If at any time the pair has a question, they raise their help card to let the teacher know they need assistance.

During CWPT sessions, the teacher moves around the classroom and can award bonus points for appropriate tutoring behavior (Maheady, Harper, & Sacca, 1988a). For example, teachers can award tutors for: (a) clearly presenting questions to tutors, (b) awarding appropriate number of points, (c) correcting errors appropriately, and (d) providing tutee with supportive comments (Maheady Harper, & Sacca, 1988a).

After each student has been tutor and tutee they total their points. Total points for each pair are calculated into team points. Teachers can announce and reinforce daily team and weekly team winners. Teachers may use praise as reinforcement or provide extrinsic rewards. See Appendix C for an overview of CWPT process. Tutoring occurs three to four times per week and is usually followed by a weekly quiz or test (Maheady, Harper & Sacca, 1988a).

**CWPT research.** Students with mild disabilities (e.g., learning disabilities, mild intellectual disabilities, behavior disorders) have historically encountered educational and interpersonal difficulties in the general education classroom (Maheady, Harper, & Mallette, 2001). However, over the past 30 years powerful interventions have been developed to meet the needs of students with mild disabilities that are educated in the general education classroom. CWPT is one of the instructional strategies that has been empirically validated for students with and without disabilities. In addition to being validated in elementary classrooms (e.g., Harper, Mallette, Maheady, Parkes, & Moore, 1993; Sideridis et al., 1997), CWPT has been shown to also be successful for students in middle and high school (e.g., Maheady, Sacca, & Harper, 2001).

Scholars have also developed variations of CWPT over the past 20 years. Special CWPT programs, such as Peer Assisted Learning Strategies (PALS), have also shown to be promising in meeting the academic and social needs of students with disabilities in general and special education classes (e.g., Fuchs, Fuchs, Mathes, & Simmons, 1997; McMaster, Fuchs, & Fuchs, 2006). However, the discussion of alternative CWPT programs go beyond the scope of this review; I am only reviewing CWPT studies that implemented the protocol developed by Delquadri, Greenwood and Stretton in 1980

(Greenwood, 1997). Additionally, because I provide a comprehensive review of the CWPT literature for students with mild disabilities as part of my methods (see Chapter 3), the purpose of the following review is to provide an overview of the CWPT literature research. It is not, however, to be used in determining the evidence based of CWPT. Some of the studies included in the following review include a) participants with more moderate or severe disabilities (e.g., autism) and/or (b) studies in which researchers incorporated additional elements into the CWPT protocol (e.g., Mastropieri et al., 2006). In Chapter 3, I provide a much more comprehensive and stringent approach to reviewing CWPT; specifically, I (a) use a systematic process to identify all CWPT studies for students with mild disabilities, (b) include only CWPT studies that strictly adhered to the Juniper Garden's protocol, and (c) use the 2013 quality indicators to determine the methodological rigor of the CWPT studies. The latter review allows me to determine whether CWPT research for students with mild disabilities is methodologically sound and evidence based. In contrast, the following provides a more general overview of CWPT the research that has been conducted in the field of special education. Both reviews focus only on CWPT studies that measured academic outcomes (e.g., neither review includes reports of studies that measure academic engagement or behavioral outcomes).

***Elementary school.*** At the elementary level, CWPT is designed to supplement instruction and replace seat work and lectures (Greenwood, 1997). The effects of CWPT have been studied at the elementary level in reading (Kamps, Barbetta, Leonard, Delquadri, 1994), spelling (Burks, 2004; Delquadri, Greenwood, Stretton & Hall, 1983), and health and safety (Utley et al., 2001).

*Reading***.** Kamps et al. (1994) used a multiple baseline design to examine the impact of CWPT compared to traditional reading instruction for students with and without autism. Researchers examined the effectiveness of CWPT on oral reading fluency and comprehension for 14 elementary students without disabilities and 3 elementary students with autism. CWPT sessions lasted 25-30 minutes three to four days a week. Results demonstrated CWPT increased reading fluency and comprehension for students with and without disabilities.

*Spelling***.** Delquadri et al. (1983) used a reversal design to examine the impact of CWPT on weekly spelling tests in a third grade classroom. Researchers examined the impact of CWPT for six students with learning disabilities and 18 students without disabilities. Results indicated that CWPT dramatically improved spelling performance for students with learning disabilities and decreased their error responses to those similar of peers without disabilities during baseline.

Burks (2004) used a single subject ABA design to examine the impact of CWPT for students with learning disabilities in spelling. Participants were three elementary students age 10 ($n$ = 2) and 11 ($n$ = 1). Results indicated CWPT increased percentages of words spelled correctly and, for one student, helped maintain higher accuracy after the intervention was withdrawn.

*Health and safety.* Utley et al. (2001) used a single subject ABAB design to examine the impact of CWPT on health and safety facts. Participants included five students with developmental disabilities. Results indicated that students' post-test scores increased when using CWPT compared to traditional instruction.

***Middle and high school.*** In middle and high school, CWPT is intended to increase students' focus on practice, skill building, and review (Greenwood, 1997). CWPT has been implemented and researched for middle and high school social studies (Maheady, Sacca, and Harper, 1988; Mastropieri, Scruggs, Spencer, & Fontana, 2003), vocabulary (Hughes & Frederick, 2006), science (Mastropieri et al., 2006), and math (Maheady, Sacca, & Harper, 1987).

*Social studies.* Mastropieri et al. (2003) used a group comparison design to examine the differences between CWPT and teacher directed guided notes; measures included pre and post-tests of reading fluency, comprehension strategies, and content tests (i.e., chapter tests, cumulative-delayed-recall tests, and a delayed-recall end-of-year final exam). Participants, who were assigned to either treatment (i.e., CWPT) or control (teacher-directed guided notes), included 16 students with mild disabilities (15 of them with learning disabilities). Results indicated that students in the CWPT group significantly outperformed students in the control group. Effect sizes for chapter tests ranged from 1.15 to 2.16.

Maheady, Sacca and Harper (1988) used single subject ABAB withdrawal design to examine the impact of CWPT for 20 students with mild disabilities. When CWPT was withdrawn, students performed more poorly on weekly social studies assessments (i.e., 20 item quizzes assessing content knowledge). When researchers implemented CWPT, student scores increased by an average of 18 points.

*Vocabulary.* Hughes and Fredrick (2006) used a multiple baseline design to examine the combined effects of CWPT and constant time delay (CTD). Participants included 3 students with learning disabilities and 15 students without disabilities in a

sixth grade language arts class. Results indicated that performance improved during treatment conditions and students maintained targeted vocabulary over time and were able to generalize the words across context.

*Science.* Mastropieri et al. (2006) used a group comparison design to examine the impact of CWPT using differentiated hands on instruction compared to traditional instruction for students with mild disabilities. Researchers developed three levels of science materials that included hands on learning activities. Participants included 213 students; 44 students were diagnosed with a disability (i.e., 37 with LD and 7 with ED). Results indicated that students who received CWPT with hands on instruction significantly outperformed traditional instruction group on the 34-item multiple choice post-tests ($p$ = .003) and the state proficiency test for science ($p$ = .014).

*Math.* Maheady et al. (2001) used a multiple baseline design to examine the impact on CWPT on academic performance in math. Participants included 28 students with mild disabilities and 63 students without disabilities in 9$^{th}$ and 10$^{th}$ grade math classes. Results indicated that the implementation of CWPT increased student scores by 20% on weekly math exams.

**Evidence base of CWPT.** CWPT has been extensively reported in special education literature as tool to support students with disabilities in increasing academic performance (e.g., Mastropieri & Scruggs, 2010). Results of the studies discussed previously suggest that CWPT has positive effects on academic achievement for students with disabilities in several content areas and across grade levels. Yet, not all of these studies discussed above applied CWPT in similar ways (i.e., did not follow the same

protocol) nor has there been a definitive review on the evidence base of CWPT for students with disabilities.

Although CEC has yet to adopt a protocol for determining EBPs for special education, other organizations have included CWPT in their evidence based reviews. For example, the Best Evidence Encyclopedia established CWPT as an EBP for students in elementary math. The WWC has not reviewed CWPT in elementary math, but found CWPT to have potentially positive effects in reading. However, neither of these organizations reviewed CWPT as an instructional strategy for improving academic outcomes for students with disabilities. Without a systematic evidence based review of CWPT for students with disabilities, researchers and educators cannot be completely confident that CWPT is effective for students with disabilities.

When determining EBPs in special education, it is important to keep in mind the inherent differences between general education and special education students: special education students need individualized instructional strategies in order to succeed academically, whereas many general education students will make academic growth even without individualized instruction. CWPT is an intervention that allows teachers to build instructional materials based on the individual needs of students in the class; teachers using CWPT can differentiate learning materials in ways to support the learning needs of students with disabilities. In addition, it is important to note that when identifying EBPs in special education, it is important to not only determine if the strategy is effective, but for *whom* the practice is effective. Best Evidence Encyclopedia and Promising Practices Network may identify CWPT as an EBP, but do not specify whether the strategy is successful for students with specific disabilities. Likewise, the research I reviewed on

CWPT covered a variety of disabilities and differing CWPT protocols. Despite the multiple studies indicating CWPT has positive effects on performance for students with disabilities, these studies may (a) not be of high quality and therefore may have biased results and/or (b) include too broad a definition of CWPT to identify which protocols of this intervention are indeed effective for students with disabilities. Therefore, the next step is to focus on clearly defining the CWPT protocol and determining whether CWPT can be considered an EBP for students for certain populations (i.e., students with mild disabilities).

**Summary**

Students with disabilities are being held to the expectations of their peers in general education. Whereas students without disabilities may succeed without the access to EBPs, students with disabilities need the most effective practices in order to meet the standards of general education (Cook & Schirmer, 2006). The idea of using the findings of multiple high quality studies to make responsible decisions regarding interventions is not new; in fact, the EBP movement originated out of the field of medicine and can be traced back to mid-19$^{th}$ century Paris and earlier.

The field of medicine began the EBP movement and it has since spread to other fields. In education, we are continuously working to bridge the research to practice gap. This review has outlined several reasons for the research to practice gap in education (e.g., the separateness of the research and practice communities, lack of meaningful professional development). EBPs are one essential component for bridging the research to practice gap in education.

Establishing EBPs in education has gained popularity as the demands for students have become more rigorous. Organizations, such as the WWC, have reviewed thousands of studies for different educational programs and interventions in order to establish EBPs. The WWC has begun to reviews interventions for students with disabilities, but often the programs reviewed are broad educational programs rather than discreet interventions. Because special education research often focuses on more discreet interventions and has populations of students with low incidence disabilities, many of the interventions used in special education may never be reviewed by the WWC. To date, however, CEC, the largest and most influential special education organization, has yet to adopt a protocol for determining EBPs in special education.

In order to establish EBPs in special education, it must be established what types of research designs can be included. Odom et al. (2005) maintained that in order for special education researchers to answer a variety of questions, a variety of research designs must be employed. However, only research designs that establish causality should be considered when establishing the effectiveness of educational interventions. When determining EBPs in the field of special education the following designs are typically considered: (a) randomized controlled trials, (b) quasi-experimental designs, and (c) single subject designs (Cook et al., 2008). In 2005, Gersten et al. (2005) and Horner et al. (2005) proposed quality indicators and standards for establishing EBPs in special education for group experimental and single subject research, respectively. In 2009, these quality indicators and standards were field tested by teams of researchers in the field of special education. Results of the field tests indicated the need for revisions to

the indicators and standards, most notably focusing on operationalizing the quality indicators for determining high quality studies.

In 2013, Cook et al., using the feedback from the 2009 review and input from 23 expert special education researchers, proposed a revised set of quality indicators and standards for establishing EBPs in special education. These quality indicators and standards expanded the work of Gersten et al. (2005) and Horner et al. (2005). Cook et al. attempted to operationalize the quality indicators and create a combined set of evidence based standards for group experimental and single subject designs. These indicators and standards have yet to be field tested in order to determine the inter-rater reliability and usability.

In the last section of this literature review, I provided an overview of CWPT procedures and research. CWPT has been documented as an effective practice for students with and without disabilities, but has yet to be established as an EBP in special education. In the next chapter of this proposal, I discuss procedures for reviewing the research literature on CWPT for students with learning disabilities using standards proposed by Cook et al (2013). In order to sufficiently review special education research, EBP standards should include protocol for reviewing single subject research. If not, many relevant and valid studies will not be included in evidence based reviews and, therefore, impact the results of determining CWPT and other interventions as EBPs. Special education, because of its inherent differences from general education, needs its own set of quality indicators and standards in order to effectively determine EBPs for students with disabilities.

# CHAPTER 3: METHODS

In this section I describe the methods for my dissertation research. Specifically, I state my research questions and describe procedures of my research including: (a) article selection, (b) coding of quality indicators, (c) calculating inter-rater reliability, and (d) EBP determination.

## Research Questions

1. What are estimates of inter-rater reliability for quality indicators used to identify sound intervention studies in special education proposed by Cook et al. (2013)?
    a. What is the inter-rater reliability for quality indicators proposed by Cook et al. (2013) for group comparison research across reviewed group comparison studies examining the effects of CWPT on students with mild disabilities?
    b. What is the inter-rater reliability for quality indicators proposed by Cook et al. (2013) for single subject research across reviewed single subject research studies examining the effects of CWPT on students with mild disabilities?
2. Does CWPT meet 2013 standards for an EBP in special education for students with mild disabilities according to standards for identifying EBPs in special education?

## Procedures

In order to determine (a) the inter-rater reliability score for the 2013 quality indicators and (b) whether CWPT can be considered an EBP when using the 2013 quality

indicators and standards for determining EBPs in special education, I used the following protocol:

1. Locate relevant CWPT research studies in which researchers: (a) examined academic impact of CWPT for students with LD, ED, ID, and/or ADHD and (b) employed research designs that reasonably establish causality (i.e., randomized control trials, quasi-experimental group comparison design, single subject design, and regression discontinuity design).
2. Code all included research studies using the 2013 quality indicators in order to identify methodologically sound studies for EBP review.
3. Determine inter-rater reliability scores for quality indicator coding.
4. Review methodologically sound studies against 2013 standards for determining whether CWPT can be considered an EBP for students with mild disabilities.

**Article selection procedures.** One purpose of this research was to determine whether CWPT, as developed by researchers at Juniper Gardens, is an EBP for students with mild disabilities. To locate relevant research studies, I first searched the Juniper Gardens website to identify CWPT research studies for students with mild disabilities. In the following, I describe (a) the initial search criteria for locating CWPT studies and (b) the inclusion criteria for studies included in the determining whether CWPT is an EBP for students with mild disabilities.

***Initial search criteria.*** After searching the Juniper Gardens website, I conducted a systematic search of Google Scholar to locate CWPT research studies conducted with students with LD, ED, ID, and/or ADHD. In order to find articles specific to the targeted

disability categories, I used combinations of the following sets of terms: (a) *classwide peer tutoring*, (b) *peer mediated instruction*, (c) *learning disabilities*, (d) *cognitive disabilities,* (e) *intellectual disabilities,* (f) *mental retardation*, (g) *emotional disabilities*, and (h) *ADHD*.

I initially included CWPT studies if in the abstract the researcher indicated participants included students with mild disabilities (i.e., LD, ID, ED, and/or ADHD) **and** either (a) the independent variable was CWPT as developed by Juniper Gardens or (b) the independent variable was referred to generally as peer tutoring (or a related term, e.g., peer assisted learning) but did not indicate the specific protocol used. For instance, I did not include studies in which researchers implemented PALS (a specific tutoring program which uses a different protocol than the protocol created by researchers at Juniper Gardens) as the independent variable, but did include studies where researchers referred to the independent variable as peer tutoring or peer mediated instruction. I later read studies in which researchers referred to peer tutoring or peer mediated instruction in their entirety to determine if they met full inclusion criteria (see subsequent section on "Inclusion criteria"). I also identified and examined reviews of CWPT research for students with LD, ID, ED and/or ADHD to locate additional studies that did not show up in Google Scholar.

For each article I found in the initial search, I created a Google Docs Excel sheet and recorded information (e.g., author, title, publication date). Next, I entered each of these studies into Google Scholar search engine (one at a time) and used Google's search tool to review additional publications that cited each original article. I scanned the abstracts of the articles in order to identify additional CWPT studies or reviews of

research for students with LD, ID, ED, and/or ADHD that met initial search criteria (see previous paragraphs) and entered the information for additional studies identified in the Google Excel Sheet.

For every article entered into Google Excel, I located the full text of each of the studies and reviews of CWPT research. I read through the reference pages of each article to locate additional CWPT studies and reviews of CWPT research. When I identified a potential article, I used Google Scholar or the University of Hawaii at Manoa library database to locate the full article or abstract and determined whether the article met initial search criteria (e.g., CWPT was the independent variable, research included students with LD, ED, ID, and/or ADHD). I entered additional studies that met initial search criteria into the Google Excel Sheet.

Next, I conducted hand searches of journals that published five or more CWPT studies or reviews of CWPT research that met my initial search criteria. Specifically I searched *Education & Treatment of Children, Remedial & Special* Education, *Learning Disabilities Research & Practice*, and *Exceptional Children*, in which I had located 12, 9, 7, and 6 respective articles from Google Scholar and reference searches. With the exception of *Learning Disabilities Research & Practice* and *Remedial & Special Education*, I conducted hand searches on each of these journals from 1980 (when CWPT was developed by Juniper Gardens) to present. I conducted a hand search for L*earning Disabilities Research & Practice* from 1999 to present because earlier publications were not available in print or electronically at University of Hawaii. Additionally, *Remedial & Special Education* was not available at University of Hawaii electronically or in print from 1980-1983, so I conducted a hand search of issues of that journal from 1984-

present. Last, I conducted a search of the *Dissertations and Theses* Database through the University of Hawaii Manoa library website. I used the same search terms as described in my initial Google Scholar search.

***Inclusion criteria.*** Out of 77 articles that met initial search criteria, 16 studies met inclusion criteria. I included studies in which the intervention was published in a refereed journal or otherwise publically available (e.g., available through the *Dissertations and Theses* database).

I included studies in which students were labeled with LD, ED, ID, ADHD, or as having a "mild disability" or "mild handicapped" (as termed in some earlier CWPT research). Additionally, I included studies in which students were labeled as having a relevant disability even if the authors did not provide information regarding how disability status was determined. For example, in some CWPT research studies, researchers described participants as having a mild disability (e.g., LD) but did not document how the participants were diagnosed with the particular disability (e.g., the student was classified as having a learning disability under IDEA, has an IEP, and was diagnosed using the discrepancy model). Additionally, some researchers described participants as having a "mild disability" or "mild handicap," but did not describe the specific diagnosis; these studies were also included. For studies in which researchers included students with and without disabilities, the study was only included if the researcher disaggregated results (i.e., results for students with and without disabilities were reported separately).

Participants in included studies must be ages 6 to 18 (i.e., school age). I included only CWPT studies that were implemented in a school setting (i.e., elementary school,

middle school, junior high, or high school) during regular school hours (e.g., after school tutoring was not included) in the EBP review. I did not include any studies that took place outside the school setting (e.g., off campus, clinical setting, or home).

I included only research studies with one of the following designs: (a) randomized control trials, (b) quasi-experimental group comparison design, (c) single subject design, or (d) regression discontinuity design. Quasi-experimental group comparison design differ from randomized control trials in that they lack random assignment. In other words, the researcher using a quasi-experimental group design assigns participants to treatment and control conditions using another criterion (e.g., in place groups). Additionally, to be included, researchers needed to (a) define CWPT as the independent variable, (b) target academic performance (e.g., words spelled correctly, post-test math scores) as a dependent variable, and (c) follow CWPT procedures outlined by Greenwood et al. (1997) (see Appendix D for student and teacher actions). Specifically, this protocol requires students to: (a) be placed into pairs, (b) take turns answering questions provided by teacher, (c) keep track of points earned for correct answers, and (d) take turns asking and answering the questions. Additionally, researchers must have indicated that teachers recorded: (a) individual points and (b) team points.

**Coding procedures for quality indicators.** According to Cook et al (2013), only studies that meet certain indicators of methodological rigor are included in an EBP review. More specifically, single subject studies must meet all 23 of the 2013 quality indicators that apply to single subject research to be used in the EBP review; group comparison studies must meet all 26 quality indicators that apply to group comparison designs. I read all articles that met inclusion criteria and coded each for presence of the

2013 quality indicators. In this section, I describe specific procedures for determining methodologically sound CWPT studies according 2013 quality indicators for group comparison studies and single subject designs (Cook et al., 2013).

All 16 studies that met inclusion criteria were single subject design studies. I employed the 2013 Coding Worksheet for Group Comparison and Single Subject Design (see Appendix E) to code each of the 16 studies along the 2013 quality indicators. The quality indicators are categorized into 8 different topic areas: (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) implementation fidelity, (f) internal validity, (g) outcome measures/dependent variables, and (h) data analysis. Cook et al.'s (2013) quality indicators combine indicators for both group comparison and single subject research studies; therefore, I did not code single subject research studies along all 31 indicators. Specifically, the eight indicators that only apply to group comparison studies (e.g., indicators related to attrition, effect size, and group assignment) were not applicable for the coding of the 16 CWPT single subject studies; studies were coded against the indicators that applied to solely single subject studies (e.g., the design provides at least three demonstrations of experimental effects at three different points in time, baseline phase includes three data points, and the single subject design controls for common threats to internal validity) and indicators that applied to both single subject research and group comparison research (e.g., indicators related to context/setting, participants, intervention agents, and description of practice).

For each study, I used the 2013 Coding Worksheet (Appendix E) to determine whether each study met each of the applicable quality indicators. As this was the initial application of the proposed quality indicators, I met with the lead author to discuss

interpretation and application in order to apply the 2013 quality indicators as intended by the CEC Workgroup. During the one-hour long meeting, we reviewed each of the indicators to ensure understanding. The meeting did not include practice coding for any of the quality indicators as one purpose of developing the 2013 quality indicators is for researchers (who understand research design and methodology) to apply these indicators without specific training (i.e., quality indicators were designed to be clearly defined and operationalized).

For each indicator a study clearly met (i.e., authors explicitly addressed in within the study), I marked "met" on the 2013 Coding Worksheet. When researchers of a particular CWPT study did not explicitly address a particular indicator, but did provide inferential information within the study that led me to determine an indicator was "met," I marked "met" in the table and provided written justification as to why I considered the indicator met. For example, Harper, Mallete, Maheady, Bentley, and Moore (1995) did not explicitly state identification procedures for students with disabilities (i.e., students were simply labeled with "mild disabilities"). However, in the methods section, Harper et al. stated students were placed within a self-contained classroom; therefore, I assumed students met IDEA criteria, as typically only students meeting IDEA would be placed in a self-contained classroom. When a study did not meet a quality indictor, I marked "not met" and provided written justification why the study did not meet that particular indicator. For example, Mortweet et al. (1999) did not provide a single subject graph that clearly represented outcome data (one of the quality indicators for data analysis); I marked "not met" and documented in the justification section authors only reported average scores. I left unmarked quality indicators that were not applicable (i.e.,

indicators for group comparison studies). I also made notes and questions in the justification section on quality indicators about which I was not clear on how to rate (e.g., is parent /teacher nomination for disability classification sufficient in determining disability status?). I specifically address these issues in Chapter 5. If a study did not meet one or more of the quality indicators, I indicated the study should not be used in the EBP review by circling "no" on the 2013 Coding Worksheet (see Appendix E for descriptions of the quality indicators).

**Inter-rater Reliability.** Dr. Lysandra Cook, an associate professor at the University of Hawaii served as my inter-rater. Dr. Lysandra Cook has experience and interest in EBPs, having published several articles on the topic (e.g., Cook, Shepherd, Cook, & Cook, 2012; Cook, Cook, Landrum, & Tankersley, 2008; Cook, Tankersley, Cook, & Landrum, 2008). Prior to this study, Dr. Cook did not have any experience or input into the quality indicators established by Cook et al. (2013). To prepare for coding, I discussed the coding protocol (as outlined in the previous section) with Dr. Cook and provided the necessary materials (i.e., 2013 Coding Worksheet and CWPT studies).

The inter-rater read 33% of studies (n = 5) that met the initial search criteria (see Kennedy, 2005, p. 120). The studies were randomly chosen; specifically, I assigned each study a number (1-16) and used an online random number generator (i.e., random.org) to generate five random numbers. Dr. Cook coded the five studies along each of the quality indicators following the same protocol as I described above (e.g., marked "met"/"not met," provided justification when indicators were not explicitly stated or not met). The inter-rater also documented questions and concerns relating to the quality indicators as

she coded (e.g., “What are the critical features for context and setting for single subject designs?”).

I used the kappa statistic to determine inter-rater reliability for (a) each individual study and (b) a combined score for all five single subject studies. Using kappa provides a measure of the magnitude of agreement and can be employed in any situation when two or more independent observers are evaluating the same thing (Viera & Garrett, 2005). Whereas other inter-rater formulas (e.g., percentage agreement) do not control for agreement by chance and thus provide inflated agreement scores, kappa controls for the chance that two observers will randomly agree. The kappa statistic indicates the degree to which the 2013 quality indicators are well defined and operationalized; the stronger the kappa score, the more confident one can have that different raters assess the presence of the 2013 quality indicators consistently. I used the following formula to compute kappa:

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Pr(a) = Observed percentage of agreement

Pre (e) = Expected percentage of agreement

I also used percentage agreement formula—(A ÷ [A + D]) x 100%, where A is agreements and D is disagreements—to determine inter-rater reliability scores for each of the 23 quality indicators separately. To do this, I counted number of agreements and disagreements for each quality indicator. Quality indicators with higher inter-rater reliability (i.e., >80%, see Horner et. al., 2005) may be considered more operationally defined; whereas quality indicators with low inter-rater scores (<80%). may indicate the

need for the CEC Workgroup to provide further explanation of the specific indicator in order to increase inter-rater reliability scores (i.e., low inter-rater reliability indicates a quality indicator may not be operationally defined enough for two independent researchers to apply consistently).

All inter-rater reliability scores (i.e., kappa and percentage agreements) and questions (from both raters) regarding the interpretation for quality indicators were documented and sent to the CEC Workgroup after initial coding was completed. In response, the CEC Workgroup made several clarifications to quality indicators (see Appendix F for questions submitted to workgroup). After these clarifications were made, I coded the remaining 11 CWPT studies.

**Evaluation procedures for determining EBPs.** Only studies that are found to be methodologically sound (i.e., meet all applicable quality indicators) can be included in an EBP review (Cook et al., 2013). For this study, all CWPT studies found to be methodologically sound were used to determine whether CWPT is an EBP. In the following section, I describe the protocol for determining whether CWPT is an EBP according to standards proposed by Cook et al. (2013).

***2013 combined standards for group experimental and single subject design*.** Standards proposed by Cook et al. (2013) categorized an intervention's evidence into the following categories: (a) EBPs, (b) potential EBP, (c) insufficient evidence, (d) mixed evidence, and (e) negative effects. For each CWPT study that I determined methodologically sound, I used the Worksheet for Determining Effects of Single Subject Studies to find whether the study had a positive, negative or neutral effect (see Appendix

G). Then, I would use the 2013 EBP Determination Worksheet (see Appendix H) to determine the status of CWPT as an EBP. After reviewing all included CWPT studies 2013 EBP standards, I would determine whether CWPT is an EBP for students with mild disabilities.

# CHAPTER 4: RESULTS

I analyzed 16 single subject research studies to determine (a) correspondence with the 2013 quality indicators and (b) the evidence based status of CWPT for students with mild disabilities. Five of the 16 studies were coded for inter-rater reliability. In the following section, I report (a) kappa scores as estimates of the inter-rater reliability of individual studies and the five studies combined and (b) percentage agreement across each of the 23 quality indicators applicable for single subject research. I also report the quality of each of the 16 studies (i.e., indicate which quality indicators were met or not met in each study). Finally, I report on the status of CWPT as an EBP.

## Inter-Rater Reliability

Five studies were randomly coded for inter-rater reliability. In total, I calculated six separate kappa scores—one for each of the five studies coded by two raters, and one for total kappa across these studies. To measure the reliability for each quality indicator separately, I calculated percentage agreement for each of 23 quality indicators for single-subject studies across the five studies.

**Inter-rater agreement for studies.** I created five kappa tables in order to document the types of agreements and disagreements raters had within each study (see Appendix I). A sixth kappa table was to document agreements and disagreements for all 5 studies combined. I used an online Kappa calculator to generate each of the Kappa scores. Kappa scores for individual studies ranged from 0.16 (SE = 0.25) to 1.0 (SE = 0.0). Kappa score for the five studies combined was 0.67 (SE = 0.06). See Table 1 for Kappa scores.

Table 1

*Kappa Statistics*

| Study | Kappa | Standard Error |
|---|---|---|
| Maheady, Harper, & Sacca (1988b) | 1.0 | 0.0 |
| Harper, Mallete, Maheady, Parkes, & Moore (1993) | 0.16 | 0.25 |
| Stevens (1998) | 0.35 | 0.35 |
| Burks (2004) | 0.74 | 0.14 |
| Bowman-Perrot, Greenwood, & Tapia (2007) | 0.83 | 0.16 |
| TOTAL Kappa | 0.64 | 0.08 |

**Inter-rater agreement for quality indicators.** Inter-rater agreement was calculated for each quality indicator across the 5 single subject studies. Specifically, I used percentage agreement (see Kennedy, 2005, p. 116) to calculate inter-rater reliability for each of the 23 quality indicators across the five single subject studies (see Table 2). Inter-rater agreement by item (i.e., quality indicator) ranged from 60% - 100%. The mean was 86.89% and both mode and median were 100%.

Table 2

*Percentage Agreement for Quality Indicators*

| Quality Indicator | Agreements | Disagreements | Percentage Agreement |
|---|---|---|---|
| **Context and Setting** | | | |
| Critical Features | 3 | 2 | 60% |
| **Participants** | | | |
| Demographics | 4 | 1 | 80% |
| Disability | 3 | 2 | 60% |
| **Intervention Agents** | | | |
| Role | 3 | 2 | 60% |
| Training | 5 | 0 | 100% |
| **Description of Practice** | | | |
| Procedures | 5 | 0 | 100% |
| Materials | 5 | 0 | 100% |
| **Implementation Fidelity** | | | |
| Adherence | 5 | 0 | 100% |

| | | | |
|---|---|---|---|
| Dosage | 5 | 0 | 100% |
| Assessment of | 4 | 1 | 80% |
| **Internal Validity** | | | |
| Researcher Control | 5 | 0 | 100% |
| Baseline Described | 5 | 0 | 100% |
| No treatment during baseline | 5 | 0 | 100% |
| Three effects | 4 | 1 | 80% |
| Three baseline data points | 3 | 2 | 60% |
| Design | 4 | 1 | 80% |
| **Outcome/Dependent Variable** | | | |
| Socially Important | 5 | 0 | 100% |
| Measurement of dv defined | 5 | 0 | 100% |
| Outcomes reported | 5 | 0 | 100% |
| Frequency of measures | 3 | 2 | 60% |
| Reliability | 4 | 1 | 80% |

| | | | |
|---|---|---|---|
| **Data Analysis** | | | |
| Linked to unit of analysis | 5 | 0 | 100% |
| Graph | 5 | 0 | 100% |

**Quality of CWPT Studies**

Out of 77 CWPT studies that met initial search criteria, 16 studies met inclusion criteria. Whereas several group comparison studies examining CWPT met initial search criteria (e.g., Scruggs & Ostguthorpe, 1986; Mastropieri, Scruggs, Spencer, & Fontana, 2003; McDuffie, Mastropieri, & Scruggs, 2009), none followed the Juniper Gardens protocol (e.g., other instructional elements were added to the CWPT protocol). Therefore, all included studies in this EBP review were single subject research designs. Across the 16 included studies, researchers employed the following single subject research designs: ABAB (*n*= 8), alternating treatment design (*n*=2), ABA (*n*=2), AB (*n*=1), BAB (*n*=1), multiple probe design (*n*=1), and multiple baseline design (n=1).

In the following, I report coding results of individual studies and evidence for why each study did not meet particular quality indicators. Only information relevant to the review of CWPT as an EBP for students with disabilities was used when coding for quality indicators. For example, if a study's participants included students with and without disabilities (e.g., Sideridis et al., 1997) only the students with disabilities are referred to and used in the context of coding for quality indicators. Additionally, if multiple outcomes were measured within a study (e.g., Stevens, 1998 measured both

student engagement and math achievement), only outcomes relevant to this review (i.e., academic outcomes) are discussed. See Table 3 for coding results for all 16 studies. I coded the remaining 11 studies (i.e., studies that were not coded for inter-rater reliability) after the CEC workgroup clarifications were made (see below) and agreement was reached for the five studies coded for inter-rater reliability. The five studies coded for inter-rater reliability are marked with an (*).

**Delquadri, Greenwood, Stretton, and Hall (1983).** Delquadri et al.'s (1983) CWPT study was the earliest publication that met inclusion criteria. Researchers used an ABAB reversal design to determine effects of CWPT on spelling achievement for six third grade children with learning disabilities. Researchers met all indicators related to: (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) outcome measures/dependent variables, and (f) data analysis. For quality indicators related to implementation fidelity, Delquadri et al. provided sufficient visual evidence (i.e., a single subject graph) to meet the quality indicator related to dosage, but did not report implementation fidelity related to adherence of CWPT protocol. Delquadri et al.'s study met five of six quality indicators related to internal validity, but because researchers collected only one data point in the return to baseline phase the study did not meet the corresponding quality indicator (i.e., baseline phases must include three data points). In total, Delquadri et al.'s study met 21 out of the 23 applicable quality indicators.

***Maheady, Harper, and Sacca (1988b).** Researchers used an ABAB reversal design across two secondary resource settings to determine the effects of CWPT on academic performance in social studies for 20 students with mild disabilities. The study

met all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) internal validity, (f) outcome measures/dependent variables, and (g) data analysis. For quality indicators related to implementation fidelity, Maheady, Harper, and Sacca (1988) provided sufficient evidence (i.e., visual data) to meet quality indicator related to dosage, but did not address implementation fidelity related to adherence of CWPT protocol. In total, Maheady, Harper, and Sacca's study met 22 out of the 23 applicable quality indicators.

**Maheady, Sacca, and Harper (1988).** Researchers used an ABAB design to determine the effects of CWPT on academic achievement in social studies for 14 tenth grade students with mild disabilities. The study met all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, and, (e) data analysis. Maheady, Sacca and Harper did provide evidence (i.e., single subject graph) to meet the implementation quality indicators related to dosage, but did not address implementation fidelity related to adherence of CWPT protocol. Maheady, Sacca and Harper's study met five out of six quality indicators related to internal validity, but because researchers collected only one data point in the return to baseline phase the study did not meet the quality indicator requiring three baseline data points. Maheady, Sacca, and Harper addressed four of five indicators related to outcome measures/dependent variables but did not report inter-observer reliability on quiz scores and, therefore, the study did not meet this particular quality indicator. In total, Maheady, Sacca and Harper's study met 20 out of the 23 applicable quality indicators.

**Bell, Young, Blair, and Nelson (1991).** Researchers used a multiple baseline design across participants to determine the effects of CWPT on academic achievement in

social studies for seven students with behavioral disabilities. Researchers addressed all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, and (e) data analysis. For quality indicators related to implementation fidelity, Bell et al. (1991) provided evidence (i.e., single subject graph) to meet the quality indicator related to dosage, but did not address implementation fidelity related to adherence of CWPT protocol. Bell et al. addressed four of six quality indicators related to internal validity. Although Bell et al.'s multiple baseline included 14 phases across seven students, the design did not delay the intervention across all seven participants. In fact, visual evidence (i.e., graphed data) illustrated only one staggered introduction of the intervention (i.e., Students 1, 2,3, and 4 received intervention simultaneously and 6 and 7 were introduced to CWPT approximately two weeks later). In turn, this design only allowed the researcher to show effects of CWPT at two different times and, therefore, researchers did not meet the quality indicator of internal validity that requires at least three demonstrations of experimental control. That is, valid multiple baseline designs require researchers to establish clear phases (including delay of intervention) across at least three participants, settings, or behaviors (Holcomb, Wolery, & Gast, 1994). Bell et al.'s (1991) study met four of five quality indicators related to outcome measures/dependent variables. Researchers did not address inter-observer reliability on quiz scores and, therefore, did not meet this quality indicator. In total, Bell et al.'s study met 19 out of 23 applicable quality indicators.

**Harper, Mallette, & Moore (1991).** Researchers used an AB design to determine the effects of CWPT on spelling achievement for 12 elementary school children with mild intellectual disabilities. Researchers addressed all quality indicators related to (a)

context and setting, (b) participants, and (c) description of practice. For quality indicators related to intervention agents, Harper et al. provided sufficient information regarding the intervention agent (i.e., the teacher) but did not provide any information to how the teacher knew how to implement CWPT within the study; as a result this study did not meet the quality indicator related to the description of training for the intervention agent. For quality indicators related to implementation fidelity, Harper et al. provided a figure that was sufficient to determine that implementation fidelity related to dosage was addressed. Researchers did not report implementation fidelity related to adherence to the CWPT protocol. Harper et al. addressed four of six quality indicators related to internal validity; however, researchers implemented an AB design and therefore did not provide at least three demonstrations of experimental effects or control for threats to internal validity. Harper et al.'s study met four of five quality indicators related to outcome measures/dependent variables; researchers did not address inter-observer reliability on scoring spelling tests and, therefore, did not meet this quality indicator. Additionally, these researchers did not include a graph that clearly represented the outcomes of the study. Specifically, the graph provided in the research study did not include baseline data. In total, Harper et al.'s study met 17 out of the 23 applicable quality indicators.

**Dupaul and Henningson (1993).** Researchers used an ABAB reversal design to determine the effects of CWPT on math achievement for a seven year old boy diagnosed with ADHD. Dupaul and Henningson (1993) addressed all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) internal validity and, (f) data analysis. Dupaul and Henningson provided visual data (i.e., single subject graph) to meet the implementation quality indicator related

to dosage, but did not report implementation fidelity related to adherence and, therefore, the study did not meet that quality indicator. Dupaul and Henningson's study met four of five quality indicators related to outcome measures/dependent variables; researchers did not address inter-observer reliability on curriculum based measures and, therefore, the study did not meet this quality indicator. In total, Dupaul and Henningson's study met 21 out of the 23 applicable quality indicators

***Harper, Mallette, Maheady, Parkes, and Moore (1993).** Researchers used a variation of an alternating treatment design to determine the effects of CWPT on spelling achievement for eight elementary students with mild disabilities (i.e., learning disabilities and mild intellectual disabilities). The study met all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) implementation fidelity, and (f) data analysis. Harper et al. (1993) addressed five of six quality indicators related to internal validity. Harper et al. used an alternating treatment design with two phases that included (a) test on words that student had not practiced (weekly pre-test) and (b) test on same words after students used CWPT to practice (weekly post-test). The research design did not meet criteria for an alternating treatment design; alternating treatment designs require a researcher employ a minimum of two interventions to one behavior that is reversible, and determine the order of implementing the independent variables (usually in random order) (Holcomb et al., 1994). Harper et al.'s alternating treatment design did not include two treatment phases; rather, it involved a no treatment pre-test and a posttest after CWPT. One purpose of researchers choosing an alternating treatment design is to determine effects of one intervention over another. Students received no instruction other than CWPT, so it is

cannot be determine whether a different type of intervention (e.g., direct instruction on subtraction problems) would have been more effective than CWPT. Thus, it was determined the design of this study did not control for common threats to internal validity. Harper et al. addressed four of five quality indicators for outcome measures; the study did not meet the quality indicators regarding evidence of inter-observer reliability. Specifically, the researchers did not report inter-rater reliability results for quiz scores. In total, Harper et al.'s study met 20 out of the 23 applicable quality indicators.

**Harper, Mallete, Maheady, Bentley, and Moore (1995).** Researchers used a variation of an alternating treatment design to determine the effects of CWPT on math achievement for eight elementary school children with mild intellectual disabilities (i.e., learning disability, $n = 4$; mild intellectual disability, $n = 3$; emotional disability, $n = 1$). Researchers addressed all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, and (d) description of practice. For quality indicators related to implementation fidelity, Harper et al. provided visual evidence (i.e., table) to meet the quality indicator related to dosage. Researchers did not address implementation fidelity related to adherence  Harper et al. met five of six quality indicators related to internal validity; researchers did not meet the quality indicator requiring researchers to employ a design that controls for threats to internal validity. Harper et al. (1995) employed a similar design found in Harper et al. (1993); and for the same reasons as previously described, did not meet criteria for an alternating treatment design. In Harper et al.'s (1995) alternating treatment design phases included (a) no treatment and (b) CWPT. Additionally, Harper et al. did not include a graph that clearly represented the outcomes of this particular study. Specifically, only a table reporting pre and post-test

scores was provided. In total, Harper et al.'s (1995) study met 19 out of the 23 applicable quality indicators.

**Matheson (1997).** The researcher used a multiple baseline design across subjects to determine effects of CWPT on spelling performance for three fourth grade students with ADHD. Matheson (1997) addressed all indicators related to (a) context and setting, (b) participants, (c) description of practice, (d) internal validity, and (e) data analysis. For quality indicators relating to intervention agents, Matheson (1997) provided sufficient information on the role of the teacher as the intervention agent; however, the researcher did not indicate how teachers were trained to conduct CWPT. For quality indicators related to implementation fidelity, Matheson provided sufficient visual evidence (i.e., single subject graph) to meet the quality indicator related to dosage, but did not report implementation fidelity related to adherence to CWPT protocol. For outcome measures/dependent variables, Matheson's study met four of five quality indicators; Matheson did not address inter-observer reliability related CBM scores. In total Matheson's study met 20 out of 23 quality indicators.

**Sideridis et al. (1997).** Researchers used an ABAB reversal design to determine the effects of CWPT on academic performance in social studies for three sixth grade students with mild disabilities enrolled in a general education classroom. Researchers addressed all indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) implementation fidelity, (f) internal validity, and (g) data analysis. Sideridis et al.'s study met all but one of the quality indicators related to outcome measures/dependent variables; the authors did not directly report inter-rater reliability on quiz scores. Researchers only reported that corrections were made when

mistakes occurred when grading. In total, Sideridis et al.'s study met 22 out of 23 applicable quality indicators.

**DuPaul, Ervin, Hook, and McGoey (1998).** Researchers used an ABAB reversal design across 18 classrooms to determine the effects of CWPT on academic performance (i.e., math or spelling test scores) for 18 elementary students with ADHD. Researchers addressed all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) implementation fidelity, and (e) outcome measures/dependent variables. DuPaul et al. met five of six quality indicators related to internal validity. Researchers did not provide three data points for baseline phases; specifically, researchers included only mean scores and ranges of pre-test scores for each of the 18 participants. DuPaul et al. addressed one of two quality indicators related to data analysis; researchers did not provide a single subject graph that represented the data collected for each of the 18 participants. In total, DuPaul et al.'s study met 21 out of 23 applicable quality indicators.

***Stevens (1998).** Researcher described using a multiple baseline and reversal design to determine the effects of CWPT on mathematics performance across two high school students with ADHD. Stevens addressed all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, and (e) data analysis. For quality indicators related to implementation fidelity, Stevens (1998) assessed implementation fidelity related to both (a) adherence and (b) dosage using a 35-item checklist; however, Stevens did not report exactly how often the researcher used the checklist to measure implementation fidelity and, therefore, the study did not meet the quality indicator related to assessing implementation fidelity regularly

throughout the intervention. It appears that Stevens's (1998) did not employ a reversal and multiple baseline design. The researcher employed an ABAB reversal designs for one student ("Joe") and an ABA design for another ("Keith"). Because an ABA design does not meet design criteria for controlling threats to internal validity, I examined only the ABAB design in order to review the quality indicators for internal validity. Stevens's study met three of six quality indicators for internal validity. The researcher did not provide sufficient information regarding curriculum used in baseline conditions. The ABAB design did not include three data points in all three baseline phases (i.e., return to baseline included only two data points). In addition, the first treatment phase of CWPT was considered a training phase in which a graduate assistant implemented CWPT. Therefore, it was determined that the ABAB design did not control for threats to internal validity, as treatment phases should remain consistent throughout the study in order to determine effects were due to treatment as opposed to other variables (i.e., trainer). Steven's study met four of five quality indicators related to the dependent variable. Stevens did not report inter-observer reliability on math CBM scoring and, therefore did not meet the corresponding quality indicator. In total, Steven's dissertation met 18 out of 23 applicable quality indicators.

**Mortweet et al. (1999).** Researchers used an ABAB reversal to determine the effects of CWPT on spelling achievement for four elementary students with mild intellectual disabilities. Mortweet et al. (1999) addressed all indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, and (e) implementation fidelity. Mortweet et al.'s study met five of six quality indicators related to internal validity. Researchers reported only average scores across both

treatment and baseline phases, and thus did not meet the quality indicator requiring researchers to show three points for each baseline phase. Mortweet et al. addressed three of five quality indicators relating to outcome measures/dependent. Researchers reported collecting data across five weeks, but because only the average scores were reported, the frequency of measures could not be confirmed with visual analysis; thus, the corresponding quality indicator was not met. In addition, the study did not meet the quality indicator related to reliability because researchers did not report inter-rater reliability on spelling quizzes. Mortweet et al.'s study met one of two quality indicators related to data analysis; researchers did not provide a single subject graph that represented the data collection for each phase. In total, Mortweet et al.'s study met 19 out of 23 applicable quality indicators.

**Utley et al. (2001).** Researchers used a BAB reversal design to determine the effects of CWPT on academic achievement in health class for five elementary students with developmental disabilities. Although this participants in this study were labeled with developmental disabilities, this study was included in the review because students had cognitive scores similar to those diagnosed with intellectual disabilities (IQ ranged from 52-69) Utley et al.'s study met all quality indicators related to (a) context and setting, (b) participants, (c) intervention agents, (d) description of practice, (e) implementation fidelity, (f) outcome measures, and g) data analysis. Utley et al. addressed three out of six indicators for internal validity. Researchers employed a BAB design, which did not allow researchers to demonstrate three experimental effects of CWPT or control for common threats to internal validity. Additionally, only two data points were collected in baseline condition. In total, Utley et al.'s study met 20 out of the 23 applicable quality indicators.

***Burks (2004).** The researcher reported using an ABA design to determine the effects of CWPT on spelling accuracy for three elementary students with learning disabilities. Burks' (2004) study met all quality indicators related to (a) context and setting, (b) participants, and (c) description of practice. For quality indicators related to intervention agents, Burks provided sufficient information regarding the role the teacher as the intervention agent, but did not indicate how teachers were trained for CWPT. Burks did not address implementation fidelity related to adherence of CWPT. The researcher provided sufficient evidence (i.e., a table) to determine that the quality indicator related to dosage was met. The study met four of six quality indicators related to internal validity. The ABA design employed did not allow the researcher to demonstrate three experimental effects or control for threats to internal validity. Burks's study met three of five indicators related to outcome measures/dependent variables; the researcher did not describe how she measured the dependent variable of spelling accuracy and did not report inter-observer agreement on scoring spelling accuracy. For quality indicators related to data analysis, Burks's (2004) study met one of two quality indicators; specifically, the researcher did not provide a visual graph displaying data points. Burks provided all data in table format. In total, Burks's study met 15 out of the 23 applicable quality indicators.

***Bowman-Perrot et al. (2007).** Only Study 1 in Bowman-Perrot et al.'s (2007) publication was used for this review; study 2 involved additional interventions combined with CWPT. Researchers reported using an ABAB reversal design to determine the effects of CWPT on spelling performance students with emotional disabilities ($n$ =19). Bowmant-Perrot et al.'s study met all quality indicators related to (a) context and setting,

(b) intervention agents, (c) description of practice, (e) implementation fidelity, (f) outcome measures/dependent variables, and (g) data analysis. For quality indicators relating to participants, Bowman-Perrot's et al.'s study met one of two quality indicators. Researchers did not provide sufficient information regarding how students were identified nor enough information about the classroom to determine the students were receiving IDEA services. Information was provided that all classes were led by special education teachers, but raters determined this information was insufficient to meet the quality indicator. The research design that was visually presented with a graph for Study 1 represented an ABA design (although an ABAB design was discussed throughout both Study 1 and 2). Therefore, this study met only four of six quality indicators for internal validity. Specifically, an ABA design does not provide three demonstrations of experimental effects or control for common threats to internal validity. In total, Bowman-Perrot's study met 19 out of 23 indicators.

Table 3

*Coding Results for Quality Indicators*

| *Quality Indicator* | *Delquadri, Greenwood, Stretton, & Hall (1983)* | *Maheady, Harper & Sacca (1988b)** | *Maheady, Sacca & Harper (1988)* | *Bell, Young, Blair & Nelson (1990)* | *Harper, Mallete, & Moore (1991)* | *DuPaul & Henningson (1993)* | *Harper, Mallete, Maheady, Parkes & Moore (1993)** | *Harper, Mallette, Maheady, Bently & Moore (1995)* |
|---|---|---|---|---|---|---|---|---|
| **Context & Setting** | | | | | | | | |
| Critical Features | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Participants** | | | | | | | | |
| Demographics | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Disability | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Intervention Agents** | | | | | | | | |
| Role | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Training | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes |
| **Description of Practice** | | | | | | | | |
| Procedures | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Materials | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Implementation Fidelity** | | | | | | | | |
| Adherence | No | No | No | No | No | No | Yes | No |
| Dosage | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Assessment of | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Internal Validity** | | | | | | | | |
| Researcher control | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| curriculum described | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| No treatment during baseline | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Three effects | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| Three baseline data points | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| Design | Yes | Yes | Yes | No | No | Yes | No | No |
| **Outcome/Dependent Variable** | | | | | | | | |
| Socially important | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Measurement of dv defined | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Outcomes reported | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Frequency of measures | Yes | Yes | Yes | Yes | Yes | Yes | No | No |
| Reliability | Yes | Yes | No | No | No | No | No | Yes |
| **Data Analysis** | | | | | | | | |
| Linked to unit of analysis | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Graph | Yes | Yes | Yes | Yes | No | Yes | Yes | No |
| **Number of Quality Indicators Met** | 21/23 | 22/23 | 20/23 | 19/23 | 17/23 | 21/23 | 20/23 | 19/23 |

***** *Indicates study coded for inter-rater reliability*

| *Quality Indicator* | *Matheson (1997)* | *Sideridis et al. (1997)* | *DuPaul, Ervin, Hook & McGoey (1998)* | *Stevens (1998)** | *Mortweet et al. (1999)* | *Utley et al. (2001)* | *Burks* (2004)* | *Bowman-Perrott, Greenwood & Tapia (2007)** |
|---|---|---|---|---|---|---|---|---|
| **Context & Setting** | | | | | | | | |
| Critical Features | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Participants** | | | | | | | | |
| Demographics | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Disability | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No |
| **Intervention Agents** | | | | | | | | |
| Role | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Training | No | Yes | Yes | Yes | Yes | Yes | No | Yes |
| **Description of Practice** | | | | | | | | |
| Procedures | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Materials | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Implementation Fidelity** | | | | | | | | |
| Adherence | No | Yes | Yes | Yes | Yes | Yes | No | Yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dosage | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Assessment of | Yes | Yes | Yes | No | Yes | Yes | Yes | No |
| **Internal Validity** | | | | | | | | |
| Researcher control | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Baseline described | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes |
| No treatment during baseline | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Three effects | Yes | Yes | Yes | Yes | Yes | No | No | No |
| Three baseline data points | Yes | Yes | No | No | No | No | Yes | Yes |
| Design | Yes | Yes | Yes | No | Yes | No | No | No |
| **Outcome/Dependent Variable** | | | | | | | | |
| Socially important | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Measurement of dv defined | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes |
| Outcomes reported | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Frequency of measures | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reliability | No | No | Yes | No | No | Yes | No | Yes |
| **Data Analysis** | | | | | | | | |
| Linked to unit of analysis | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Graph | Yes | Yes | No | Yes | No | Yes | No | Yes |
| **Number of Quality Indicators Met** | 20/23 | 22/23 | 21/23 | 18/23 | 19/23 | 20/23 | 15/23 | 19/23 |

***** *Indicates study coded for inter-rater reliability*

**Evidence Based Status of CWPT**

Cook et al. required that a study must meet all of the quality indicators applicable to the study's design. Therefore none of the studies could be used in determining whether CWPT is an EBP for students with mild disabilities and CWPT should be classified as a practice with insufficient evidence (see Appendix H for description).

# CHAPTER 5: DISCUSSION

The purpose of this study was to determine (a) the inter-rater reliability scores for Cook et al.'s (2013) quality indicators for determining EBPs in special education and (b) whether CWPT can be considered an EBP for students with mild disabilities using these quality indicators and standards. Five out of 16 single subject studies were scored for inter-rater reliability. Kappa statistics for individual studies ranged from k = 0.16 to k = 1.0; the five studies combined yielded k = 0.64. %. For individual quality indicators, inter-rater reliability was perfect (100%) for the majority of quality indicators (13/23), moderate (80%) for five quality indicators, and low (60%) for five quality indicators. None of the 16 studies met criteria for being methodologically sound (i.e., met all applicable quality indicators); therefore CWPT cannot be considered an EBP according to the 2013 quality indicators and standards. The following sections discuss (a) limitations of the study, (b) interpretation of findings for the two research questions presented, and (c) implications of this study for research and practice.

## Limitations

This study is not without its limitations, which should be considered when interpreting the findings. First, I included only studies that strictly adhered to the CWPT protocol developed by Juniper Gardens. Therefore, studies that implemented CWPT using (a) an alternative protocol (e.g., eliminated any steps outlined by Juniper Gardens) or (b) CWPT in combination with another strategy or intervention were not included for review. Results may have been different if I used broader inclusion criteria that identified more studies involving CWPT implemented using slightly different procedures. Therefore, results pertaining to the quality indicators of CWPT studies and the EBP status

of CWPT can only be discussed in the context of the 16 studies included in this dissertation and not generalized to other CWPT research.

Additionally, this initial field test of the 2013 quality indicators and standards was conducted before any publication(s) became available to help guide my review process. Thus, I did not have the same material that will likely be available for future field tests of the 2013 quality indicators and my interpretation of the quality indicators may have been correspondingly impaired. Instead, I used the list of quality indicators and standards and correspondence with the chair of the CEC Workgroup to guide coding the CWPT studies. Moreover, coding for quality indicators of studies involves some subjective judgment. My ratings may be inappropriately low or high (though inter-rater reliability did show an acceptable rate of non-chance agreement between my ratings and the ratings of a second rater).

CEC Workgroup clarifications, which were presented after my initial coding and inter-rater reliability measures were completed, assisted in further defining quality indicators and allowed me to more effectively code the remaining 11 studies. For the five studies coded by both raters, I made no changes to quality indicator coding after CEC Workgroup clarifications, with the exception of finding agreement on quality indicators on which raters disagreed. Therefore, coding guidelines differed slightly between the 5 studies scored for inter-rater reliability and the 11 studies coded after CEC Workgroup clarifications. For example, both reviewers coded all quality indicators related to implementation fidelity as "met" for Harper et al. (1993) although the authors did not explicitly report an official implementation fidelity score. The coding of Harper et al.'s study was not changed even though CEC workgroup clarifications suggested that inter-

rater reliability scores must be explicitly reported. Thus, outcomes for the quality of the five studies may be slightly inflated for some items and possibly under-rated for other items.

**Interpretation of Findings**

In this section I discuss and interpret (a) inter-rater reliability scores for quality indicators, (b) the overall quality of CWPT research, and (c) how the results of this study relate to the extant literature and research of EBPs in the field of special education.

**Reliability of 2013 quality indicators.** The CEC Workgroup proposed quality indicators essential for determining methodologically sound intervention studies in order to allow special education researchers to determine which studies have the minimal methodological features to have confidence in their findings (Cook et al., 2013). Hence, research question #1 (i.e., what are estimates of inter-rater reliability for the 2013 set of quality indicators used to identify EBPs in special education?) is important; special education researchers must be able to apply these quality indicators reliably for consistent and valid determination of study quality and trustworthiness. Although study quality is a generally important consideration for research consumers, it plays a particularly important role in evidence-based reviews. Studies are only included in EBP reviews if they are considered methodologically sound (i.e., meet all quality indicators).

Research question #1 has two sub-questions: What is the inter-rater reliability for quality indicators proposed by Cook et al. (2013) for (a) group comparison research studies examining the effects of CWPT on students with mild disabilities and (b) single subject research studies examining the effects of CWPT on students with mild disabilities? The former sub-question could not be addressed through this initial field test

of the 2013 quality indicators, as there were no CWPT group comparison studies that met inclusion criteria. Therefore, in the following, I discuss the results of inter-rater reliability scores for the five single subject studies coded for inter-rater reliability.

***Interpretation of kappa scores.*** Kappa statistics for the five individual CWPT single subject studies ranged from k = 0.16 to 1.0. According to benchmarks for kappa scores suggested by Landis and Koch's (1977), one study was considered to have slight level agreement (i.e., k = 0.16; Harper et al., 1993);one study was considered to have a fair level of agreement (i.e., k= 0.35; Stevens, 1998); one study was considered to have substantial agreement (i.e., k = 0.74, Burks, 2004); and two studies were in the almost perfect level of agreement; (i.e., k= 0.83, 1.0; Bowman-Perrot et al., 2007; Maheady, Harper, & Sacca,1988b, respectively). The kappa scores across all five studies was k = 0.643; which is considered a substantial strength of agreement (Landis & Koch, 1977).

Inter-rater reliability for the five CWPT studies seems promising; quality indicators developed by the CEC workgroup were operationalized to the extent that the two raters coding five studies found acceptable agreement on their presence. However, there was variability in reliability between studies and percentage agreement for each of the 23 applicable quality indicators (i.e., quality indicators that apply to only single subject studies and quality indicators that apply to both single subject and group design studies) also showed variability regarding the reliability of quality indicators.

***Interpretation of percentage agreement.*** Percentage of agreement for each of the quality indicators ranged from 60% to 100%. Inter-rater reliability was perfect (100%) for the majority of quality indicators (13/23), moderate (80%) for five quality indicators, and low (60%) for five quality indicators. Items that had low inter-rater scores (i.e., 60%)

included quality indicators relating to (a) describing characteristics of critical features of the context or setting, (b) describing the disability or risk status of the participant, (c) describing the role of the intervention agent, (d) baseline phases containing three data points, and (e) frequency and timing of the outcome measures.

All inter-rater reliability scores (i.e., kappa and percentage agreements) and questions (from both raters) regarding the interpretation for quality indicators were documented and sent to the CEC Workgroup after initial coding was completed. Several questions concerning the interpretation of quality indicators submitted to CEC workgroup after the initial coding related directly to indicators with the lowest inter-rater scores (e.g., what is acceptable when determining risk status; do alternating treatment designs need to have a baseline phase?). Raters had more difficulty coding quality indicators with low inter-rater reliability (i.e., 60%), suggesting the quality indicators with low inter-rater reliability needed further clarification. See Appendix F for questions submitted to CEC Workgroup.

In response to questions submitted by both raters, the CEC Workgroup made several clarifications and minor changes to the quality indicators. Responses from the Workgroup assisted raters in finding agreement on quality indicators for which there was initial disagreement and clarified my understanding when coding the additional 11 studies included in the EBP review. In the following, I discuss (a) each quality indicator with an inter-rater reliability score less than 100%, (b) CEC Workgroup clarifications and/or changes to the particular quality indicators, and (c) final agreements reached for each quality indicator on which raters initially disagreed. Appendix J (a) shows each

rater's coding on each quality indicator that was disagreed upon, (b) rater justification on their particular rating, and (c) final determination for a study's quality indicator.

*Characteristics of context or setting.* Inter-rater agreement for the quality indicator requiring the researcher to describe the characteristics of critical features of context or setting was 60%. For both studies in which there was disagreement (i.e., Burks, 2004; Stevens, 1998), one rater coded both studies as "met", whereas the other rater coded the quality indicator for both studies as "not met". The rater who coded both studies as "not met" indicated there was not enough information regarding the actual classroom, student/teacher ratio, and diversity. The CEC Workgroup clarified that the primary importance of this quality indicator is to establish whether a study falls within the parameter of the review (Cook et al., 2013). Specifically, because I only included CWPT studies that were conducted in a school setting, that was the only evidence required to meet this quality indicator. Therefore, final determination was both studies met this particular quality indicator.

*Participant demographics.* Inter-rater agreement for the quality indicator requiring researchers to describe participant demographics was 80%. One rater noted Harper et al. (1993) provided sufficient information on gender, disability, and age, but did not provide information on socioeconomic status or language and therefore did not meet this quality indicator. Similar to the quality indicator related to context and setting, the CEC Workgroup clarified that to meet the quality indicator researchers need to describe characteristics of participants in order to determine whether the study should be included in the EBP review. Harper et al. provided critical information for the CWPT

review for students with mild disabilities; therefore, raters determined Harper et al.'s study met this particular quality indicator.

*Disability or risk status of participants.* Inter-rater agreement for the quality indicator requiring researchers to describe the disability or risk status of participants was 60%. For both studies on which there was disagreement (i.e., Burks, 2004; Harper et al., 1993), one rater coded both studies "met" whereas the other rater coded the quality indicator for both studies "not met." The rater who coded both studies as "not met" indicated that there was not enough information regarding how participants were diagnosed with the particular disability. The other rater noted that although the researchers in both studies did not explicitly state how participants were diagnosed with a disability, placement in a resource room (Burks, 2004) and self-contained classroom (Harper et al., 1993) was sufficient to determine students were diagnosed with a disability under IDEA. The CEC Workgroup clarified that (a) documentation of disability category and (b) placement in a special education setting is acceptable for meeting this quality indicator. Therefore, raters made the final determination that both Burks's (2004) and Harper et al.'s (1993) studies met the quality indicator related to describing the disability status of participants.

*Role and background of the intervention agent.* Inter-rater agreement for the quality indicator requiring researchers to describe the role and background of the intervention agent was 60%. For both studies in which there was disagreement (i.e., Burks, 2004; Stevens, 1998), one rater assumed both studies met the quality indicator whereas the other rater coded the quality indicators as not met. The latter rater noted that there was not enough information regarding the backgrounds of the intervention agent;

the only description provided in both studies was that teachers were the interventionists. Similar to the quality indicators related to context and setting and participants, the CEC Workgroup specified to meet the quality indicator that (a) enough information should be provided to determine the interventionist of the study and (b) interventionists are described in enough detail to determine whether the study should be included in the EBP review (Cook et al., 2013). For the purpose of this EBP review, raters agreed that it was clear in both Burk's (2004) and Stevens (1998) who conducted the intervention. Additionally, I had no additional inclusion criteria related to interventionist, therefore, raters made the final determination that both Burks's and Steven's studies met the quality indicator related to describing role of the intervention agent.

*Regular assessment of implementation fidelity.* Inter-rater agreement for the quality indicator requiring the researcher to assess implementation fidelity regularly throughout the study was 80%. Bowman-Perrot et al. (2007) implemented an ABA design over 14 weeks and assessed implementation fidelity one time during the semester. The CEC Workgroup provided clarifications that implementation fidelity related to adherence must be reported (i.e., researchers must report a specific level of adherence), and further explained that implementation fidelity should be reported across phases. After discussion, both researchers agreed Bowman-Perrot et al.'s (2007) study did not meet this quality indicator.

*Research design provides three demonstrations of experimental control.* Inter-rater agreement for the quality indicator requiring single subject researchers to provide three or more demonstrations of experimental control was 80%. In Stevens' (1998) first treatment phase a graduate assistant facilitated CWPT; this phase was described as the

training phase. In the second treatment phase, the teacher implemented CWPT. The disagreement on this indicator was not a result of poor operationalization of the quality indicator, but due to each rater's interpretation of the design. One rater commented the design was actually an ABAC (i.e., the first treatment phase was a training phase led by a graduate assistant, the second treatment phase was led by the classroom teacher); however the other rater argued that the researcher had the *opportunity* to demonstrate experimental control with the phases presented. After discussion, it was decided that the issue regarding Stevens' design was more applicable to the quality indicator requiring the researcher to employ a design that controls for common threats to internal validity. Therefore, raters determined Stevens met this particular quality indicator.

*Baseline phase includes three data points.* Inter-rater agreement for the quality indicator related to requiring single subject researchers to include three data points in baseline phases was 60%. Stevens (1998) collected three data points in initial baseline and two data points in the second baseline. One rater agreed that this was appropriate; some experts recommend less than three data points at return to baseline is acceptable. For example, Kennedy (2005) suggests that baseline needs to be long enough to sample a pattern; therefore one rater assumed that because a three point baseline was established initially, the return to baseline did not need three points (i.e., the return to baseline phase showed significant decrease in academic performance). The CEC Workgroup clarified that all baseline phases should have a minimum of three data points unless there is justification by author for reasons such as (a) measuring severe and/or dangerous problem behaviors and/or (b) zero baseline behaviors with no likelihood of improvement without

intervention. Raters agreed that Stevens' study did not meet any of the exceptions and, therefore, did not meet this quality indicator.

Harper et al. (1993) used a variation of the alternating treatment design. One rater determined that because an alternating treatment design was employed, the researchers did not need a baseline phase; the other rater noted that because the researchers included only one data point for each phase, the study did not meet this quality indicator. CEC Workgroup clarified that for alternating treatment designs a baseline phase is not necessary. Therefore the raters agreed that Harper et al.'s study met this quality indicator.

*Design controls for common threats to internal validity.* Inter-rater agreement for the quality indicator requiring single subject researchers to control for common threats to internal validity was 80%. As described in a previous section, in Stevens' (1998) first treatment phase a graduate assistant facilitated CWPT; during the second treatment phase, the teacher implemented CWPT. CWPT protocol was the same during both treatment phases. One rater agreed that this design controlled for threats to internal validity, but the other rater suggested that because the two treatment phases were implemented by different interventionists this may introduce a threat. Rater discussion led to the decision that Stevens did not meet this quality indicator; specifically because researchers employed only two treatment conditions and each involved a different interventionist. Therefore, although the reversal design allowed for three possible demonstrations of effect, the demonstrations of effect were in regard to two different treatments (CWPT training conducted by researchers, CWPT conducted by teachers).

*Frequency and timing of outcome measures are appropriate.* Inter-rater agreement for the quality indicator requiring appropriate frequency and timing of

outcome measures was 60%. Unlike other quality indicators with 60% reliability, coding for these studies was inconsistent between raters. For Stevens' (1998) study, one rater coded the study as "not met" because researchers collected only two data points in the second baseline. The other rater coded the quality indicator as "met" because some experts (e.g., Kennedy, 2005) suggest that baseline data needs to only be long enough to sample a pattern; the rater determined that the dramatic decrease in return to baseline was sufficient. Similar to the quality indicator requiring baseline phases have three data points, unless there is justification, this item requires both treatment and intervention phases of a single subject study to have three data points per phase. Therefore, raters determined that Stevens did not meet this particular quality indicator.

Harper et al. (1993) reported employing a variant of an alternating treatment design and, therefore, included only one data point per phase. Both raters agreed that Harper et al. did not meet several quality indicators related to internal validity because the design did not meet the criteria for an alternating treatment design. Hence, one rater coded this study as "not met" because researchers included only one data point per phase. The other rater noted that this study met the quality indicator because researchers used an alternating treatment design (and therefore baseline was not required). Because both raters agreed this design did not meet criteria for an alternating treatment design, they also determined the study did not include the appropriate number of data points per phase (i.e., three). Therefore Harper et al. did not meet this quality indicator.

*Adequate evidence of internal reliability is described.* Inter-rater agreement for the quality indicator requiring researchers to provide adequate evidence of reliability (i.e., internal, inter-observer, test-re-test, and/or parallel form) was 80%. In Harper et al.'s

(1993) study, one rater coded this study "not met" and indicated that, although the researchers reported that all tests were scored by a second scorer, researchers did not report a reliability score. The CEC Workgroup confirmed that it is unacceptable to only report that inter-observer reliability was assessed; researchers must report a specific level of adherence. Therefore, raters made the final determination that Harper et al. did not meet this quality indicator.

***Interpretations of coding disagreements.*** Eight out of 10 quality indicators on which raters disagreed were specifically addressed and clarified by the CEC Workgroup (i.e., quality indicators related to context and setting, participant demographics, disability status, role of the interventionist, baseline phases, timing of outcome measures, frequency of implementation fidelity, and evidence of internal reliability), making final coding determinations for these quality indicators rather straightforward. Disagreements on quality indicators related to the design (a) showing at least three experimental effects and (b) controlling for threats to internal validity seem to be based on discrepancy between how the researcher defined the design (e.g., ABAB) and the way the design was actually implemented within the study. Specifically, Stevens (1998) described using a multiple baseline design across two participants. Had Stevens clearly implemented a multiple baseline and included visual data (i.e., single subject graph) to show this type of design, the raters would have been able to clearly code the quality indicator requiring the design to provide experimental effects at three different points in time as "not met." Multiple baseline designs require a minimum of three participants or three phases (Kennedy, 2005). However, Stevens' visual data showed that an AB design was used for one participant and for the other participant raters had difficulty deciding whether the design

met criteria for an ABAB or an ABAC. Further, six of 13 total disagreement occurred for Stevens' study, which was a dissertation. One possible reason for these discrepancies may be because, although dissertation research is publicly available, it does not go through a peer review process for publication in a journal. Therefore, clarity of writing and design flaws may have led to higher rates of disagreements between raters.

Further, it seems possible that when a quality indicator is clearly "present" or completely "absent" coding is rather straightforward. However, it becomes more difficult to determine whether a study meets a particular quality indicator when the quality indicator is "partially accomplished" (i.e., there is some evidence of its presence, but authors do not explain in sufficient detail for indication of absolute presence on a dichotomous scale). For example, raters disagreed on whether Stevens (1998) provided evidence of the opportunity to show three demonstrations of experimental effects. Although Stevens reported implementing a multiple baseline design, both raters used visual data to determine that the design used did not reflect the requirements of a multiple baseline (e.g., baselines established concurrently, independent variable sequentially introduced across participants).

However, one rater believed that Stevens used an ABAB design and, thus, met the criteria for three demonstrations of effect. The second rater, however, concluded the design was an ABAC because the two treatment phases were led by different interventionists (i.e., trainer, teacher). The researcher's description of the design led to confusion when coding this particular quality indicator. This type of reporting leads to a more subjective determination ("how much information is enough?"). Thus, it seems that research studies with moderate quality or moderate reporting may lead to lower inter-

rater agreement. On the other hand, studies that completely leave out information regarding a quality indicators (i.e., studies with low methodological rigor) or studies that clearly include and report information on quality indicators (i.e., high methodological studies) may yield higher inter-rater reliability scores.

Several quality indicators had perfect inter-rater reliability (e.g., quality indicators related to (a) training of intervention agents, (b) intervention procedures, and (c) researcher control and manipulating the independent variable). This may be due to differences in particular quality indicators; specifically, coding some quality indicators may be more straightforward (i.e., easier to determine a presence or absence) than coding other quality indicators. For example, quality indicators requiring evidence of (a) relevant materials, (b) reporting all outcome measures, and (d) a single subject graph all earned perfect inter-rater agreement (100%) and seem to be easily coded using a dichotomous scale (i.e., raters can clearly justify its presence or absence); whereas quality with lower inter-rater agreement may not be as easily coded using a dichotomous scale. For example, quality indicators related to (a) frequency and timing of outcome measures, (b) evidence of role of intervention agent, and (c) evidence of disability, all with inter-rater agreement of 60%, may not have been as straightforward when raters were interpreting how much information is "enough" for meeting the quality indicators.

Harper et al.'s (1993) study had the second highest number of disagreements (i.e., raters disagreed on 5 out of 23 quality indicators) and earned the lowest kappa statistic ($k = 0.16$). Unlike Stevens' (1998) dissertation study, all of the quality indicators on which there were disagreements were addressed by clarifications from the CEC Workgroup. Similarly, the three disagreements in Burks' (2004) study were all addressed

with clarifications from the CEC Workgroup. Although further field tests of the revised quality indicators need to be conducted, the CEC Workgroup clarified all of the quality indicators with unacceptable inter-rater agreement (i.e., 60%), which may lead to more reliable coding and higher inter-rater agreement on the aforementioned items when future coding is conducted with the guidelines generated from this pilot.

***Interpretation of reliability scores in light of literature and theory.*** In 2005, Gersten et al. and Horner et al. presented sets of quality indicators and standards for determining EBPs for the field of special education with the hope that the quality indicators would be "field-tested and refined, then considered useful by journal editors and reviewers of federal grant proposals" (Gersten et al., 2005, p. 150). Since their publication, several researchers have used the quality indicators to describe studies included in meta-analyses (e.g., Bellini & Akullian, 2007; Browder, Spooner, Ahlgrim-Delzell, Harris, & Wakeman, 2008; Jitendra, DuPaul, Someki, & Tresco, 2008; Flippin, Reszka, & Watson, 2010; Jitendra, Burgess & Garcia, 2011) and when designing and conducting their research (e.g., Hume & Odom, 2007; Lane et al., 2008). The work conducted by Browder et al. (2006) and evidence-based reviews in the 2009 special issue of *Exceptional Children* (i.e., Baker et al., 2009; Browder et al., 2009; Chard et al., 2009; Lane et al., 2009; and Montague & Dietz, 2009) provided extensive feedback and recommendations in order to refine the 2005 quality indicators. Specifically, these EBP reviews were crucial in developing the 2013 quality indicators and standards presented by the CEC Workgroup. In the following, I discuss how the 2013 quality indicators may improve both the reliability and validity in comparison to previous quality indicators (i.e., Gersten et al., 2005; Horner et al., 2005).

*Reliability of quality indicators.* Because Gersten et al. (2005) and Horner et al. (2005) did not specify the methods to be used for coding quality indicators for single subject and group experimental research studies, researchers conducting field tests using the 2005 quality indicators applied various methods when evaluating the methodological rigor of an intervention study (e.g., some have devised rating scales, others have used a dichotomous met/not met approach). Hence, inter-rater agreement scores could not be synthesized to broadly examine the reliability of the 2005 quality indicators. In addition, preliminary findings suggested that the rating approaches used may have had low inter-rater reliability (Cook et al., 2009). Determining reliability is an essential step in adopting a set of quality indicators for determining methodologically sound studies; specifically, it is important to understand how consistently quality indicators are coded among different researchers reviewing the same intervention studies.

The CEC Workgroup specified that the 2013 quality indicators to be coded "met" or "not met" in order to streamline the efforts of coding procedures and determine a more accurate means of reporting reliability. This dissertation research serves as the initial field test of the 2013 quality indicators and results indicate promising reliability statistics. In fact, kappa statistics for the 5 single subject research studies reviewed by both raters was $k = 0.64$, which indicates substantial agreement (Landis & Koch, 1977). Further, after inter-rater reliability scores and questions were presented to the CEC workgroup, quality indicators were further clarified and defined (e.g., alternating baselines do not require a baseline phase; specific level of reliability scores must be reported). Changes made to the quality indicators allowed raters to easily come to final agreement, which suggests that future field tests may lead to even stronger reliability of the quality indicators.

Reviewers from the field test of the 2005 quality indicators concluded that the many of the quality indicators needed to be further operationalized in order to code for their presence. For example, Lane et al. (2009) further defined the Horner et al.'s (2005) quality indicator requiring measurement of implementation fidelity but added that implementation fidelity must also be recorded. In contrast with field tests of the 2005 quality indicators, we did not alter or refine definitions of quality indicators before coding CWPT intervention studies for the 2013 quality indicators. Further, percentage agreement for 2013 individual quality indicators ranged from 60% to 100%, with 13 quality indicators having 100% inter-rater reliability. Although inter-rater reliability ranges were lower than those reported for both Browder et al. (2009) and Lane et al. (2009), it is important to note that we double coded only 5 studies, whereas Browder et al. and Lane et al. double coded 10 and 16 studies respectively.

*Validity of quality indicators.* Baker et al. (2009) emphasized the importance of establishing validity of the quality indicators used to determine the methodological rigor of a study. Validity, in terms of the quality indicators, involves determining how well the quality indicators actually define the construct of a methodologically sound study. In other words, in order for the quality indicators to be considered valid, it is essential to determine that the quality indicators include the necessary components of a methodologically sound study and eliminate unnecessary elements. The work by Gersten et al. (2005) and Horner et al. (2005) was instrumental in presenting an initial set of quality indicators for special education research and, because these indicators were developed by top experts in the field, may be assumed to have content validity (i.e., the experts are knowledgeable and, therefore, included essential components of

methodologically sound studies when developing quality indicators). However, the 2005 quality indicators were never empirically examined to determine their validity.

Cook et al.'s (2013) Delphi study partially addresses Baker et al.'s (2009) recommendation to ensure the validity of the quality indicators. Twenty-four expert special education researchers were asked to rate on a 1 to 4 scale (where 1 is strongly disagree and 4 was strongly agree) each area of quality indicators. Results of this qualitative study (as discussed in detail in Chapter 2) imply that the quality indicators developed by the CEC Workgroup reflect a consensus regarding the critical aspects of a methodologically sound study in special education research. For the 2013 quality indicators, the Delphi study was a step in examining the content validity, which was not directly assessed in the development of the 2005 quality indicators. However, the validity of the final 2013 quality should undergo further empirical validation related to criterion validity (i.e., examine the relation of quality indicators to a study's effect). In the field of medicine, Juni, Witschi, Bloch, & Egger (1999) applied 25 different quality indicator scales to 17 clinical trials comparing treatments for postoperative thrombosis. Researchers found that using different scales produced different conclusions regarding the benefit of treatment. Specifically, studies identified as methodologically sound differed according to which scale was used and the methodologically sound studies identified (for different scales) reached different conclusions regarding the best treatment options. It will be important for the field of special education to investigate whether (a) effect sizes of studies categorized as sound and not sound on the basis of the 2013 quality indicators differ and (b) effect size is related to the presence of particular quality indicators.

**Evaluation of the evidence base of CWPT.** None of the 16 CWPT single subject research studies met all 23 of the quality indicators applicable for single subject research design studies (i.e., indicators that apply specifically to single subject research and indicators that apply to both single subject and group design). Although the quality of each study varied, all 16 CWPT studies met 11 of the 23 quality indicators. Specifically, all studies addressed (a) characteristics of the critical context or setting relevant to the review, (b) participant demographics, (c) role and background of the intervention agent, (d) detailed intervention procedures, e) relevant intervention materials, (f) implementation fidelity related to dosage, g) researcher control over the independent variable, (h) non-access to treatment intervention during baseline/comparison phases, (i) socially important outcomes, (j) reporting all outcome measures, and (k) data analysis techniques were appropriately linked to unit of analysis. Fifteen of the 16 studies also met quality indicators related to (a) describing disability status, (b) describing baseline, and (c) defining the dependent variable. Additionally, with the exception of three studies (i.e., Burks, 2004; Harper et al., 1991; Matheson, 1997), studies sufficiently described training procedures for intervention agents.

That said, with the exception of Sideridis et al. (1997) and Maheady, Harper, and Sacca (1988b), all studies failed to meet one or more quality indicators in multiple methodological categories. Ten studies did not provide evidence of reliability for outcome measures, making this particular quality indicator the most often not addressed. Although two studies inferred that they conducted reliability measures (i.e., Harper et al., 1993; Sideridis et al., 1997) for outcome measures, they did not report a reliability score and, thus, did not meet the particular quality indicator. However, it is possible that this

particular quality indicator was not applied appropriately or applied too strictly. This quality indicator requires a description of adequate evidence of internal reliability, inter-observer reliability, test-retest reliability, and/or parallel form reliability, as relevant (Cook et al., 2013). While coding for this particular quality indicator, I required evidence of inter-rater reliability of outcome measures. Because all of the dependent variables in the 16 studies were some type of academic outcome, all outcome measures included some type of performance on a test, quiz, or CBM. Unlike measures of classroom behavior (e.g., time on-task), which require researchers to clearly define in order to identify the observable behavior, the academic outcomes targeted in this review were much more straightforward in terms of assessing accurately (e.g., number of words spelled correctly, accuracy of subtraction problems). This may have been a reason for researchers not consistently reporting inter-rater reliability on these measures.

In addition, all outcome measures included in the review were teacher made, and none of the studies directly reported internal reliability of measures. To meet the quality indicator, only one type of reliability must be measured. Whereas several studies did report inter-rater reliability, the internal reliability of the measures is left unknown. Thus, a research study that reports adequate inter-rater reliability on spelling tests can be assumed to have reliable scoring procedures, but it cannot be determined that there was internal consistency reliability (i.e., the degree to which different test items probe the same construct and have similar results); therefore, even studies that met this particular quality indicator may actually have problems with other forms of reliability (e.g., internal consistency). It could be argued that the test items for the CWPT studies have face validity (i.e., outcomes are directly linked to what is being taught) and therefore, do not

need to report internal reliability of items. However, it remains that the studies reviewed would have more methodological rigor had authors directly measured (and reported) evidence of internal reliability.

With the exception of implementation fidelity related to dosage, quality indicators related to implementation fidelity were also met infrequently. According CEC Workgroup clarifications (Cook et al., 2013), researchers using single subject design do not need to specifically address implementation fidelity related to dosage if they (a) provide a graph that clearly shows the number of intervention sessions and (b) state duration of intervention sessions. All studies provided sufficient information to meet this particular quality indicator. Seven of the 16 CWPT studies met the quality indicator on implementation fidelity related to adherence. The results of this quality indicator seem related to the date of publication. Implementation fidelity seems to be a quality indicator frequently overlooked in past decades; Gresham, MacMillan, Beebe-Frankenberger, and Bocian (2000) found that only 12 out of 65 (18.5%) intervention studies that were published in three major learning disabilities journals from 1995-1999 measured implementation fidelity. In this EBP review, with the exception of Matheson (1997) and Burks (2004), studies published between 1997-2007 specifically addressed implementation fidelity related to adherence; whereas studies published between 1983-1995 (with the exception of Harper et al., 1993) did not meet the quality indicator. It should also be noted that Harper et al. (1993) reported assessing implementation fidelity but did not report a specific level of adherence; thus, had coding been changed after CEC Workgroup clarifications were presented, the study would not have met the quality indicator.

It is concerning that that several CWPT studies did not meet quality indicators related to internal validity. Eight out of 16 studies did not demonstrate adequate use of a single subject design in order to control for threats of internal validity. In four studies (i.e., Burks, 2004; Bowman-Perrot et al., 2007; Harper et al., 1991; Utley, 2001) researchers employed AB, ABA, or BAB designs. These designs do not allow the researcher to demonstrate three experimental effects of CWPT or control for threats to internal validity. Further, four other studies (i.e., Bell et al., 1990; Harper et al., 1993; Harper et al., 1995; Stevens; 1998) did not employ the single subject research designs appropriately and, thus, did not meet internal validity criteria. For example, Harper et al. (1993) and Harper et al. (1995) implemented a variation of the alternating treatment design. However, the design did not provide alternating treatments; they simply measured the difference of the effects of no treatment to the effects of CWPT on spelling performance, which does not appropriately implement the alternating treatment design. Researchers in this study may have used these designs because of the appeal (i.e., ease of implementation) to the practitioners they were working with. Smith et al. (2013) suggests that practitioners are not as concerned with ruling out all possible explanations for student achievement, but rather focus on if an intervention can be implemented in their own classroom and if it works for their own students. Although the designs in these studies do not demonstrate experimental control or establish causality, the results indicated a positive change in academic achievement. In addition, data collection procedures seemed to align with what a teacher would engage in on a typical day (e.g., spelling tests, health content quizzes) making this intervention practical for practitioners. Thus, it seems that the next step for these researchers would be to use a single subject

design that controls for threats of internal validity while still maintaining its usefulness and ease of implementation for practitioners.

Five studies did not meet the quality indicator requiring at least three data points for each baseline phase. The most common way studies did not meet this particular quality indicator was that researchers provided only one point in return to baseline (i.e., Delquadri et al., 1983; Maheady, Sacca, & Harper, 1988b; Stevens, 1998). Researchers may have included only one data point in return to baseline because of the dramatic change in level from intervention to baseline phase. However, unless the researcher establishes reason for only one baseline data point (e.g., measuring severe or dangerous behavior; zero baseline behavior with no likelihood of improvement) all baseline phases must include a minimum of three data points. None of the CWPT studies met the exceptions to the criteria. In single subject research it is important to establish trends across each phase in order to document experimental control. When researchers only provide one data point in a phase, it does not provide enough evidence of functional control (e.g., one data point does not establish a pattern of intervention effects). Alternatively, two studies included only average scores for baseline measures (Mortweet et al., 1997; DuPaul et al., 1998), which also goes against the principles of single subject design. In order to use visual analysis to determine an intervention's effect on a dependent variable, it is essential that all data points be included and reported visually.

One of the key components of single subject research is the use of visual analysis to determine an intervention's effect on the dependent variable. Surprisingly, four of the 16 studies did not provide a graph to allow the reviewer to analyze the data collected. Specifically, Harper et al. (1991) only provided visual data of intervention phases;

baseline data were reported in a separate table. Burks (2004), DuPaul et al. (1988), and Mortweet et al. (1999) presented all data in table form and, therefore, visual analysis was not possible. It may be possible that these researchers believed that reporting data in tables was a sufficient means in reporting results and that a research consumer may use the information reported to graph and visually analyze the data points. However, using visual analysis for making decisions about introduction and withdrawal of the treatment is an inherent principle of single subject research. In other words, researchers should collect and analyze visual data as they are conducting their research. It is also critical for research consumers to have an accurate graph of the study's data in order to meaningfully examine the elements of visual analysis (e.g., variability, trend, overlap).

Overall, Sideridis et al. (1997) and Maheady, Harper, and Sacca (1998b) provided sufficient information to meet quality indicators in seven of the eight categories. Sideridis et al. was the only CWPT study that met all but one of the quality indicators. Six studies (i.e., Bowman-Perrot et al., 2007; Delquadri et al., 1983; DuPaul & Henningson, DuPaul et al., 1998; Stevens, 1998; Utley et al., 2001) met all quality indicators in six out of the eight categories. Five studies (i.e., Bell et al., 1990; Harper et al., 1993; Maheady, Sacca, & Harper, 1988b; Matheson, 1997; Mortweet et al., 1999) met all quality indicators in five categories. The remaining three studies provided sufficient evidence in four or fewer categories. In order to determine CWPT as an EBP, the practice would need to be supported by at least five methodologically sound studies with positive effects and at least 20 participants. None of the studies in this review met the methodological rigor necessary to be included in the EBP review.

One possible reason for these findings is that Cook et al. (2013) set too rigorous a standard to be expected for educational research. Although the 2013 quality indicators are indeed rigorous it remains that they are representative of research methods that are widely accepted in the field. As Chard et al. (2009) argued, quality indicators "serve as clear targets that researchers should consider both in designing their research as well as in describing the results of their research for dissemination" (p.277). Future researchers should plan, conduct, and report their research using the 2013 quality indicators as a guide so that sound studies will be available to conclusively determine the efficacy of CWPT.

The question, then, remains whether the current CWPT research can be used as a meaningful indicator of the intervention's effectiveness for students with mild disabilities. Whereas previous reviews of CWPT research (e.g., Byrd, 1990; Greenwood, 1997; Ryan, Reid, Epstein, 2004; Stenhoff & Lignugaris/Kraft, 2007) have suggested that the intervention is effective for students with disabilities, only Stenhoff and Lignugaris/Kraft (2007) included an examination of the methodological rigor of the studies included for review (yet these authors did not exclude any study in which quality indicators were not present—which was over half of the studies included). Additionally, whereas previous reviews included studies with various peer tutoring models (e.g., cross age peer tutoring or models where tutors and tutees did not switch roles), this dissertation study limited the scope of the research to only CWPT studies that employed the Juniper Gardens protocol. Similar to previous reviews of CWPT, results of individual studies in the current EBP review present evidence of the effectiveness of CWPT; however because none of the studies met standards of a methodologically sound study, effects should be

interpreted with caution. Nonetheless, this does not suggest the current CWPT research is meaningless or the intervention should be eliminated for consideration as an educational practice. Results do warrant (a) caution when interpreting the effects of CWPT studies and (b) further (high quality) research to be conducted in order to be more confident in its potential as an EBP.

**Implications for Research and Practice**

I am hesitant to draw firm conclusions regarding the use of CWPT for students with mild disabilities based on the findings of this review. Similar to EBP reviews for repeated readings, function based interventions, and cognitive strategy instruction for mathematics conducted in 2009 (i.e., Chard et al.; Lane et al.; Montague & Dietz), CWPT tutoring was not found to be an EBP for students with mild disabilities. However, similar to the aforementioned interventions, CWPT was founded from the several well-established theoretical frameworks (i.e., effective teaching theory, eco-behavior analysis, and social learning; as cited in Maheady & Gard, 2010). Additionally, the majority of the studies reviewed ($n$=9) examined the impact of CWPT on spelling performance for elementary students with disabilities and all reported an increase of correctly spelled words on weekly tests. Thus, while it is important to remain tentative when determining the effectiveness of CWPT, it seems this intervention may be promising for elementary students with disabilities in the area of spelling.

In contrast, the impact of CWPT on mathematics performance was not as promising. Three studies (i.e., DuPaul and Henningson, 1993; Harper et al., 1995; and DuPaul et al., 1998) examined the impact of CWPT on mathematics performance for elementary students with mild disabilities. Whereas Harper et al. and DuPaul et al.

reported CWPT to positive impact academic achievement, DuPaul and Henningson reported minimal effects of CWPT on CBM math probes. Similarly, Stevens (1998) found CWPT to have minimal impact on the mathematics performance of two secondary students.

Overall, I suggest that CWPT should be used when no relevant EBP can be identified as many of the studies reviewed did report positive impacts on academic achievement, especially in the area of spelling. However, its use, as with any instructional intervention, CWPT must be used in combination with progress monitoring in order to assess its effectiveness on student achievement. Progress monitoring when using CWPT is especially important, in that, it has not been identified as an EBP for students with mild disabilities.

Although none of the studies included in this evidence based review met the criteria for being methodologically sound, the majority of studies reviewed found that CWPT tutoring significantly increased academic achievement for students with disabilities; whereas only three studies did not find CWPT to have a significant impact on academic achievement. Currently, CWPT could be considered a researched based practice (i.e., the practice has evidence of some research support but the research does not meet the methodological rigor of an EBP; see Cook & Cook, 2013). In order to provide enough methodologically sound studies to truly determine the evidence base of CWPT, I encourage researchers to conduct replication studies using the 2013 quality indicators when designing and reporting the research. CWPT studies included in this review found CWPT to have positive effects on several academic outcomes for students with mild disabilities. It seems that CWPT is a potential EBP that needs more methodologically

sound research to validate its effectiveness. Additionally, as Chard et al. (2009) recommended, it would also benefit the field if funding agencies provides replication competitions and encouraged research collaborations that would result in replicating studies in multiple regions of the country.

When conducting future research on CWPT or any intervention in special education, it will be important for researchers to use the 2013 quality indicators when planning, conducting, and writing their research study. Moreover, it is important for researchers to not only conduct systematic research using the quality indicators as a guideline, but to clearly report the methods they used. As Moyer and Finney (2005) suggested, incomplete or unclear reporting interferes with a reviewer's ability to accurately judge the methodological quality of a study. Further, it may be necessary for journals to provide web based links in order to include information regarding methods and data analysis that may not be included within the journal publication due to lack of space (Baker et al., 2009). In this EBP review, I found that quality indicators relating to reporting (a) reliability of outcome measures, (b) internal validity, and (c) three data points per phase were met with lowest incidence. In light of the findings, future researchers will need to increase the frequency with which they assess and report inter-rater agreement and internal reliability of outcome measures. In addition, when using single subject designs, researchers should take care to choose a design that controls for common threats to internal validity (e.g., ABAB, multiple baseline design, alternating treatment design) and document sufficient data points across all phases (i.e., a minimum of three).

When conducting evidence based reviews, it will be important for future reviewers to clearly define both the intervention and who the intervention is for (Browder et al., 2006). This study identified CWPT in terms that allowed only certain CWPT studies to be included (i.e., studies that implemented the protocol established by Juniper Gardens). In addition, parameters were set to include only studies involving students with mild disabilities and academic outcomes. I did not set inclusion/exclusion criteria for the age of the participants in order to maintain enough studies to conduct the EBP review. Had I chosen to narrow my inclusion criteria further (e.g., include only students with learning disabilities or include only secondary students with mild disabilities), I would not have located enough studies to determine the evidence based status of CWPT, even if all were determined methodologically sound. That being said, it is important the parameters are carefully set in order to provide enough information regarding (a) the critical components of the intervention and (b) for whom the intervention may be effective for. In Montague and Dietz's (2009) review of cognitive strategy instruction, reviewers were unable to clearly identify the major components of the strategy in review because they included many types of cognitive strategy instruction. In other words, when conducting an EBP review it is important to set inclusion parameters that are educationally meaningful. Researchers will need to balance between specificity and breadth. It is not realistic or practical to investigate whether a broad practice (e.g., function based interventions) is effective for everyone and for all outcomes. On the other hand, it is not realistic to be too specific (e.g., is CWPT effective for spelling outcomes for fifth grade boys with autism?). In special education, EBP reviews should be

conducted on a practice that can be operationalized for a particular population and outcome area (Cook & Cook, 2013).

Additional field tests for the 2013 quality indicators are also warranted and recommended. Inter-rater reliability scores from this initial field test indicate that the 2013 quality indicators can be applied with adequate inter-rater reliability. However, because the CEC Workgroup made clarifications to certain quality indicators after initial coding and reliability was completed, it will be necessary for other researchers to establish inter-rater reliability for the final set of quality indicators. Moreover, replicating this pilot study with a larger sample of studies that includes group comparison studies and different types of outcome measures will be important.

As Baker et al. (2009) suggested it is important that measurement tools used for determining methodologically sound studies be not only reliable, but also valid. Cook et al.'s (2013) Delphi study indicated that the quality indicators and standards developed by the CEC Workgroup demonstrated adequate content validity. However, further research on the empirical validity of the quality indicators related to criterion validity is warranted. The CEC Workgroup's approach to developing the quality indicators and standards set a high bar and, in turn, will exclude many studies, especially those conducted before these quality indicators were developed (Cook et al., 2009). In turn, it will be important to justify the exclusion of studies that do not meet the quality indicators. I propose two ways in which research in this area could be conducted. First, it will be important to investigate whether the presence of specific quality indicators are related to study effects (i.e., if there is no relation between the presence of a quality indicator and study effects, one can infer that the quality indicator is not differentiating between studies with poor and high

methodological rigor). Second, it will be important to investigate whether effects differ for studies that do and do not meet criteria for being methodologically sound (i.e., if there is no difference in effects, then there is no reason to exclude any studies within an EBP review).

Also important is educating future teachers on becoming quality consumers of research. The purpose of establishing EBPs in special education is to help bridge the research to practice gap. However, establishing EBPs is not sufficient to close the gap. I recommend that teacher preparation programs not only provide pre-service teachers with information and training regarding "what works" in special education, but also that special educators learn (a) how to interpret the quality of research that is available for a specific intervention and (b) the importance of reading research to keep up with innovations in the field. Thus, it will also be important for special education researchers to publish practitioner friendly papers that highlight the latest findings in special education research. It is also important for organizations (e.g., CEC) to compile an online database that is user-friendly and allows teachers to read the latest research on frequently used interventions.

**Conclusions**

This review served as the initial field test of the 2013 quality indicators and standards for establishing EBPs in the field of special education. The work conducted by Gersten et al. (2005) and Horner et al. (2005), in combination with the field tests of the 2005 quality indicators, were essential in the development of the 2013 quality indicators and standards presented by the CEC Workgroup. Although further field tests are warranted, preliminary analyses indicate the 2013 quality indicators demonstrate

acceptable reliability when reviewing single-case studies. Additional research on the empirical validity of the quality indicators and standards will need to be conducted.

In terms of the status of CWPT, I recommend that researchers conduct high quality, experimental studies—using the 2013 quality indicators as a guide when designing, implementing, and reporting the studies—to conclusively determine the effectiveness of this practice for learners with mild disabilities. Currently, the literature base lacks enough methodologically sound studies to determine the effectiveness of CWPT for students with mild disabilities. Nonetheless, CWPT is an intervention with considerable theoretical and empirical support that should not be disregarded or considered ineffective. It is important for researchers to work towards establishing and reporting effective research practices, in order to provide guidelines for conducting sound studies that can be used to meaningfully classify the effectiveness of practices. This study demonstrated that the 2013 quality indicators can be reliably applied to classify the methodological soundness of intervention studies in special education, which is an important step in identifying what really works for students with disabilities. Adopting and using the 2013 quality indicators to establish EBPs in special education will provide trustworthy evidence about what works; which, in turn, will improve the ability of special education stakeholders to provide effective and appropriate education for students with disabilities.

Appendix A
Teams and Partners Chart

Teams and Partners Chart

Week: _____5-6_____

Subject: Science

Team_____Bears_____ Team_____Packers_____

| Partners | | Partners | |
|---|---|---|---|
| **Move** | **Stay** | **Move** | **Stay** |
| John | Tierney | Cory | Jackie |
| Sara | Natalie | Jenn | Andrew |
| Derek | Jared | Sarena | Gia |
| Lola | Stephanie | Tiffany | Jeremy |
| | | | |
| | | | |
| | | | |
| | | | |

(adapted from Greenwood, Delquadri, Carta, 1997, p.10)

Appendix B

Tutoring Point sheet

Tutoring Point Sheet

Student_____ Date_____

Times through 1 2 3 4 5

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
| 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 |
| 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 |

**Bonus Points!**

(Greenwood, Delquadri, Carta, 1997, p.23)

# Appendix C

## CPWT Protocol

Step 1: Move to tutoring positions.

a. Get the attention of all students
b. Review the Move/Stay procedure
c. Tell "movers" to stand.
d. Give the students the signal to move to their partners
e. Students move to their partners

Step 2. Get ready for tutoring.

a. Pass out materials
    a. Weekly tutoring material
    b. Tutoring point sheets
    c. Help signs

b. Set time and give signal to begin tutoring

Step 3. CWPT Session

Students should write down total points for each student at the end of the session

Step 4. Clean up, report points, winning team

(Greenwood, Delquadri, Carta, 1997, p.33)

# Appendix D

## Teacher and Student Actions

| **Teacher Actions** | |
|---|---|
| 1. Tutoring Pairs | Students switch partners each week. Teachers are responsible for pairing students by random pairing or skill pairing. Random Pairing is putting students together by chance. Skill pairing refers to pairing students with similar skills or having students with higher skills work with students with lower skill sets. |
| 2. Weekly Teams | Teachers are responsible to assign students to teams. |
| 3. Movers/Stayers | Teachers post chart to remind students: (a) partnerships, (b) weekly teams, and (c) movers and stayers (see Figure 1 for Team Chart) |
| 4. Reserving Time | Tutoring sessions should last approximately 30-35 minutes |
| 5. Pre-test/Posttest design | Pretests should be given before CWPT is implemented. Posttests should be given on the content taught during CWPT. |
| 6. Teaches CWPT to students | Teacher instructs students on all CWPT protocol (Figure 3) and how to use tutoring point sheet (Figure 2) |
| 7. Provides CWPT material | Provides students with tutoring material (e.g., spelling words, math problems, science vocabulary) |
| 8. Tracks points | Teacher tracks student and team points |
| 9. Bonus Points | Teacher may add bonus points for appropriate student-student interaction or observed improvement in academic skills |
| **Student Actions** | |
| 1. Tutor/Tutee | Students take turn being tutor/tutee during each tutoring session |
| 2. Record points | Tutor should keep track of points tutee earned |

# Appendix E
## 2013 Coding Worksheet for Group Comparison and Single Subject Design

| Essential Quality Indicators | Met | Not Met | Justification |
|---|---|---|---|
| Context & Setting | -- | -- | -- |
| *Sufficient information is provided regarding the critical features of the contexts or settings relevant to the study (i.e., in which the intervention and control/ comparison/ baseline conditions occurred).* | -- | -- | -- |
| 1. Characteristics of the critical features of the context(s) or setting(s) relevant to the review (e.g., type of program[s]/classroom[s], type of school [e.g., public, private, charter, preschool], curriculum used, geographical location[s], community setting[s], socio-economic status, physical layout[s]) are described. [B] | | | |
| Participants | -- | -- | -- |
| *Sufficient information is provided to identify the population of participants to which results may be generalized and to determine/confirm whether the participants demonstrated the disability(ies) or difficulty(ies) of focus.* | -- | -- | -- |
| 1. Participant demographics relevant to the review (e.g., gender, age/grade, race/ethnicity, socio-economic status, language status) are described. [B] | | | |
| 2. Disability or risk status of participants (e.g., specific learning disability, autism spectrum disorder, behavior problem, at-risk for reading failure) and method for determining status (e.g., identified by school using state IDEA criteria, teacher nomination, standardized intelligence scale, curriculum-based measurement probes, rating scale) are described. [B] | | | |
| Intervention Agents | -- | -- | -- |
| *Sufficient information is provided regarding the critical features of the intervention agents.* | -- | -- | -- |
| 1. Role (e.g., teacher, researcher, paraprofessional, parent, volunteer, peer tutor, sibling, technological device/computer) and background variables of intervention agent(s) relevant to the review (e.g., gender, race/ethnicity, educational background/licensure, professional experience, experience with intervention) are described. [B] | | | |
| 2. If specific training (e.g., amount of training, training to a criterion) or qualifications (e.g., professional credential) are required to implement the intervention, they are described and achieved by interventionist(s). (If specific training or qualification are not required, check box) [B] | | | |
| Description of Practice | -- | -- | -- |
| *Sufficient information is provided regarding the critical features of the practice (intervention), such that the practice is clearly understood and can be reasonably replicated.* | -- | -- | -- |
| 1. Detailed intervention procedures (e.g., intervention components, instructional behaviors, critical or active elements, manualized or scripted | | | |

| | | | |
|---|---|---|---|
| procedures, dosage) and intervention agents' actions (e.g., prompts, verbalizations, physical behaviors, proximity) are described, or one or more accessible sources are cited that provide this information. [B] | | | |
| 2. When relevant, materials (e.g., manipulatives, worksheets, timers, cues, toys) are described, or one or more accessible sources are cited that provide this information. [B] | | | |
| Implementation Fidelity | -- | -- | -- |
| *The practice is implemented with fidelity.* | -- | -- | -- |
| 1. Implementation fidelity related to adherence is assessed using direct, reliable measures (e.g., observations using a checklist of critical elements of the practice) and reported. [B] | | | |
| 2. Implementation fidelity related to dosage or exposure is assessed using direct, reliable measures (e.g., observations or self-report of the duration, frequency, and/or curriculum coverage of implementation) and reported. [B] | | | |
| 3. Implementation fidelity is (a) assessed regularly throughout implementation of the intervention (e.g., beginning, middle, end of the intervention period) as appropriate for the study being conducted; (b) assessed for each interventionist, each setting, and each treatment group (or participant or other unit of analysis in single subject-research) as relevant, and (c) reported. If neither adherence (item #1) nor dosage (item #2) is assessed and reported, this item is not applicable. If either adherence (item #1) or dosage (item #2) is assessed and reported (but not both), this item applies to the type of fidelity assessed. [B] | | | |
| Internal Validity | -- | -- | -- |
| *Independent variable is under the control of experimenter.* | | | |
| 1. The researcher(s) controls and systematically manipulates the independent variable. [B] | | | |
| *The nature of services provided in control/comparison condition(s)/phases is described.* | -- | -- | -- |
| 2. The curriculum, instruction, and interventions used in control/comparison condition(s) (in group comparison studies) or baseline/comparison phases (in single-subject studies) are described (e.g., definition, duration, length, frequency, learner:instructor ratio). [B] | | | |
| 3. Access to the treatment intervention by control/comparison group(s) (in group comparison studies) or during baseline/comparison phases (in single-subject studies) is not provided or is extremely limited. [B] | | | |
| *The research design provides sufficient evidence that the independent variable causes change in the dependent variable(s).* | -- | -- | -- |
| 4. Assignment to groups is clearly and adequately described. [G] | | | |
| 5. Participants (or classrooms, schools, or other unit of analysis) are assigned to groups in one of the following ways: (a) randomly; (b) nonrandomly, but the comparison group(s) is matched very closely to the intervention group (e.g., matched on prior test scores, demographics); (c) non-randomly, but techniques are used to measure and, if meaningful differences (e.g., statistically significant difference, difference of > 0.05 | | | |

| | | | |
|---|---|---|---|
| pooled SDs for matched studies [What Works Clearinghouse, 2011]) are identified, statistically control for any differences between groups on relevant pre-test score and/or demographic characteristics (e.g., statistically adjust for confounding variable through techniques such as ANCOVA or propensity score analysis); or (d) non-randomly on the basis of a reasonable cutoff point when regression discontinuity design is used. [G] | | | |
| 6. The design provides at least three demonstrations of experimental effects at three different points in time. [S] | | | |
| 7 For single-subject research designs that use a baseline phase (alternating treatment designs do not require a baseline), all baseline phases include at least three data points (except when fewer are justified by study authors due to reasons such as [a] measuring severe or dangerous problem behaviors and [b] zero baseline behaviors with no likelihood of improvement without intervention) and establish a pattern that predicts undesirable future performance (e.g., increasing trend in problem behavior, consistently infrequent exhibition of appropriate behavior, highly variable behavior). [S] | | | |
| 8. The design controls for common threats to internal validity (e.g., ambiguous temporal precedence, history, maturation, diffusion) such that plausible, alternative explanations for findings can be reasonably ruled out. Commonly accepted designs such as reversal (ABAB), multiple baseline, changing criterion, and alternating treatment address this quality indicator when properly designed and executed, although other approaches can be accepted if the researcher(s) justifies how they rule out alternative explanation for findings/control for common threats to internal validity. [S] | | | |
| *Participants stayed with the study, so attrition is not a significant threat to internal validity.* | -- | -- | -- |
| 9. Overall attrition is low across groups (e.g., ≤ 20% in a one-year study). [G] | | | |
| 10. Differential attrition (between groups) is low (e.g., within 20% of each other) or is controlled for by adjusting for non-completers (e.g., conducting intent-to-treat analysis). [G] | | | |
| Outcome Measures/Dependent Variables | -- | -- | -- |
| *Outcome measures are applied appropriately to gauge the effect of the practice on study outcomes* | -- | -- | -- |
| 1. The outcome(s) is socially important (e.g., it constitutes or is theoretically or empirically linked to improved quality of life, an important developmental/ learning outcome, or both). [B] | | | |
| 2. Measurement of the dependent variable(s) is clearly defined and described. [B] | | | |
| 3. The effects of the intervention on all measures of the outcome(s) targeted by the review are reported (*p* levels and effect sizes [or data from which effect sizes can be calculated] for group comparison studies; graphed data for single-subject studies), not just those for which a positive effect is found. [B] | | | |
| 4. Frequency and timing of outcome measures are appropriate. For group comparison studies, outcomes must be measured at both pre- and post-test at a minimum. For single-subject studies, a minimum of 3 data points per phase must be measured (except when fewer are justified by study authors due to reasons such as [a] measuring severe and/or dangerous problem behaviors and [b] zero baseline behaviors with no likelihood of improvement without intervention). [B] | | | |

| | | | |
|---|---|---|---|
| *Outcome measures demonstrate adequate psychometrics.* | -- | -- | -- |
| 5. Adequate evidence of internal reliability, inter-observer reliability, test-retest reliability, and/or parallel form reliability, as relevant, is described (e.g., score reliability coefficient $\geq$ .80, IOA $\geq$ 80%, or Kappa $\geq$ 60%). [B] | | | |
| 6. Adequate evidence of concurrent, content, construct, or predictive validity is described (e.g., a specific validity coefficient is reported). [G] | | | |
| 7. Evidence of reliability and validity (with the exception of inter-observer reliability, which must be evaluated using data within the study) are empirically evaluated based on (a) data generated within the study (i.e., researchers use their own data) or (b) data from another study. If evidence is imported from another study, the sample and scores are similar enough to make generalization to the current study sensible. [G] | | | |
| Data Analysis | -- | -- | -- |
| *Data analysis is conducted appropriately.* | -- | -- | -- |
| 1. Data analysis techniques are appropriately linked to the unit of analysis in the study. For example, if classrooms are randomly assigned to conditions in a group comparison study, then classroom (not individual) should be the unit of analysis (with the exception of multilevel analyses such as HLM, in which multiple units of analysis exist). Similarly, if the research question for a single-subject study is stated in terms of the effect of an intervention on a classroom, then classroom-level data should be analyzed. [B] | | | |
| 2. Data analysis techniques are appropriate for comparing change in performance of two or more groups (e.g., t-tests, (M)ANOVAs, (M)ANCOVAs, hierarchical linear modeling, structural equation modeling). If atypical procedures are used, a rationale justifying the data analysis techniques is provided. [G] | | | |
| 3. A single-subject graph clearly representing outcome data across all study phases is provided for each unit of analysis (e.g., individual, classroom or other group of individuals) so that reviewers can determine the effects of the practice. Regardless of whether study authors include their own visual or other analyses of data, graphs depicting all relevant dependent variables targeted by the review should be clear enough for reviewers to draw basic conclusions about experimental control using traditional visual analysis techniques (i.e., analysis of mean, level, trend, overlap, and consistency of data patterns across phases). [S] | | | |
| *Information on effect size is reported.* | -- | -- | -- |
| 4. One or more appropriate effect size statistics (e.g., Cohen's d, Hedge's G, Glass's Δ, eta-squared) is reported for each primary outcome, even if the outcome is not statistically significant; or data are provided from which appropriate effect sizes can be calculated. [G] | | | |

**Use of Study in EBP Review**

| Label | Criteria | YES | NO* |
|---|---|---|---|
| Methodologically Sound | Meets all quality indicators | | |

# Appendix F

## Questions submitted to CEC Workgroup

- What is considered sufficient evidence for descriptions of (a) context(s)/setting(s), (b) participants, and (c) interventionist(s)
- What is the appropriate method for determining disability/risk status
- When is training required? When is training adequately achieved?
- What is appropriate for describing manipulatives?
- Is reporting of implementation fidelity as acceptable enough for meeting quality indicator?
- How do you code the quality indicator of: regularly assessing fidelity across interventionists/conditions/phases if implementation fidelity related to adherence and dosage is not addressed?
- Are baseline phases required in all designs (even alternating treatment)?
- What constitutes attrition?
- Any participants initially participating for whom data was not analyzed for any reason constitutes attrition.
- How to code for multiple outcome measures in a study, in which, some meet a quality indicator, but others do not?
- What is sufficient evidence for reliability?
- Do single subject studies require active assessment dosage of single-subject studies?
- Does baseline phases always require three data points?
- How many data points are necessary per phase?

# Appendix G

## Worksheet for Determining EBPs

| High Quality Experimental Studies | Type of Effect | High Quality Quasi-Experimental Studies | Type of Effect | High Quality Single Subject | Type of Effect |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |
| | | | | | |

| | | | |
|---|---|---|---|
| **Total # of High Quality Studies with positive effects** | | # of Participants | |
| **Ratio positive effect: neutral effect** | | # of Researchers | |
| **# of studies with Negative effects** | | # of Locations | |

| EBP criteria: | Potential EBP criteria | Insufficient Evidence | Mixed Evidence | Negative Effects |
|---|---|---|---|---|
| ☐ 2 high quality experimental studies (positive effects) OR<br>☐ 4 quasi-experimental studies (positive effects) OR<br>☐ 5 single subject studies with positive effects (and meet Test et al. (2011) criteria OR<br>☐ 50% of criteria for two or more studies above<br>☐ No high quality studies with negative effects<br>☐ 3:1 ratio of high to neutral effects | ☐ 1 high quality group experimental studies (with positive effects) OR<br>☐ 2 high quality quasi-experimental study OR<br>☐ 2 high quality single subject studies (with positive effects)<br>☐ 50% of criteria for two or more studies above<br>☐ No high quality studies with negative effects<br>☐ 2:1 ratio of positive to neutral | ☐ Insufficient number of high quality studies to meet criteria of potential EBP AND<br>☐ 1 or more high quality studies with negative effects | ☐ Meet criteria for an EBP or a potential EBP AND<br>☐ have one or more high quality studies with negative effects, but the high quality studies with negative effects do not outnumber high quality studies with positive effects OR<br>☐ has a ratio of high quality studies with to high quality studies with neutral effects is less than 2:1 | ☐ 1 or more high quality studies with negative effects<br>☐ Number of high quality studies with negative |
| EBP: Y N | Potential EBP: Y N | Insufficient Evidence: Y N | Mixed Evidence: Y N | Negative Effects: Y N |

# Appendix H

## Kappa Tables

Maheady, Harper, & Sacca (1988b):

| | Yes | No | Total |
|---|---|---|---|
| Yes | 19 | 0 | 19 |
| No | 0 | 4 | 4 |
| Total | 19 | 4 | 23 |

Burks (2004):

| | Yes | No | Total |
|---|---|---|---|
| Yes | 11 | 2 | 13 |
| No | 1 | 9 | 10 |
| Total | 12 | 11 | 23 |

Harper et al. (1993):

| | Yes | No | Total |
|---|---|---|---|
| Yes | 17 | 3 | 20 |
| No | 2 | 1 | 3 |
| Total | 19 | 4 | 23 |

Bowman-Perrot (2007)

| | Yes | No | Total |
|---|---|---|---|
| Yes | 19 | 1 | 20 |
| No | 0 | 3 | 3 |
| Total | 19 | 4 | 23 |

Stevens (1998)

| | Yes | No | Total |
|---|---|---|---|
| Yes | 14 | 5 | 19 |
| No | 1 | 4 | 4 |
| Total | 15 | 9 | 23 |

Combined Kappa

| | Yes | No | Total |
|---|---|---|---|
| Yes | 80 | 11 | 91 |
| No | 4 | 20 | 24 |
| Total | 85 | 30 | 115 |

Appendix I

Inter-rater Disagreements and Final Determination for Quality Indicators

| Quality Indicator | Study | Rater A | Rater B | Determination |
|---|---|---|---|---|
| Critical Feature of context and setting | Stevens (1998) | *Met* | *Not met*<br>2 regular ed classrooms in rural public high school in NY; no information on classrooms eg. Teacher student ratio; diversity | *Met* |
| | Burks (2004) | *Met* | *Not met*<br>LD resource room, # of students and teachers noted (p.302) | *Met* |
| Participant demographics | Harper et al. (1993) | *Met* | *Not met*<br>Gender, disability, and age noted (p.27); nothing on SES, ethnicity or language | *Met* |
| Disability | Burks (2004) | *Met*<br>Resource room placement assumes IDEA services | *Not met*<br>"received sped, diagnosed with LD" (p.302) | *Met* |
| Disability | Harper et al. (1993) | "Met"<br>Self contained classroom assumes IDEA services | "Not Met"<br>Noted as LD or MMR and as being in an SDC, but not how they were identified (p.27) | *Met* |
| Role of Interventionist | Stevens (1998) | *Not Met*<br>No background on teacher was provided | *Met*<br>Teacher became the intervention facilitator p.14 | *Met* |
| Role of Interventionist | Burks, 2004 | "Not met"<br>Only information provided was two special | "Met"<br>Not directly stated but can be inferred from procedures…(p.302) | *Met* |

| | | | | |
|---|---|---|---|---|
| | | education teachers | | |
| Implementation fidelity assessed regularly | Bowman-Perrot et al. (2007) | *Met* | *Not met*<br>Once in a semester for study 2 | *Not met* |
| Three demonstrations of effect | Stevens (1998) | *Met*<br>ABAB across one participant. Training phase was same as implementation except implemented by grad student | *Not met*<br>Supposed to be a multiple baseline, reversal design across 2 participants (which isn't 3). For one you have an ABAB, but I think technically you have an ABAC design-as the first treatment phase is training. But if you consider training and treatment equivalent (they did deliver CWPT in training, it was done by grad student) | *Met* |
| Baseline phase has three data points | Stevens (1998) | *Met*<br>First baseline phase has three points | *Not met*<br>Second baseline phase has two points | *Not Met* |
| Baseline phase has three data points | Harper et al. (1993) | *Not met*<br>One point per phase | *Met*<br>No baseline which is appropriate for alternating treatment design | *Met* |
| The design controls for common threats to internal validity | Stevens. (1998) | *Met*<br>ABAB | *Not met*<br>For one participant, but not if you consider it an ABAC | *Not met* |
| Three data points per phase | Stevens (1998) | *Met* | *Not met*<br>Only 2 points in 2nd baseline, no justification of why this may have occured | *Not met* |

| | | | | |
|---|---|---|---|---|
| Three data points per phase | Harper et al. (1993) | *Not met*<br>One data point per phase; no clear phases exist. | *Met* | *Met* |
| Adequate evidence of reliability | Harper et al. (1993) | *Met* | *Not met*<br>Notes that all tests scored by a 2nd scorer to assure accuracy, but no mention of agreement rate. | *Not met* |

# References

Arreaga-Mayer, C. (1998). Increasing active student responding and improving academic performance through classwide peer tutoring. *Intervention in School and Clinic, 34,* 89-94.

Baker, S.K., Chard, D. J., Ketterlin-Geller, L. R., Apichatabutra, C., Doabler, C. (2009). Teaching writing to at-risk students: The quality of evidence for self-regulated strategy development. *Exceptional Children, 75*, 303-220.

Bell, K., Young, R., Blair, M., Nelson, R. (1990). Facilitating mainstreaming of students with behavioral disorders using classwide peer tutoring. *School Psychology Review*, 16, 564-573.

Bellini, S. and Akullian, J. (2007). A meta-analysis of video modeling and video self-modeling interventions for children and adolescents with autism spectrum disorders. *Exceptional Children, 73*, 264-287.

Best Evidence Encyclopedia (BEE) (n.d.). *Best Evidence Encyclopedia.* Johns Hopkins University. Retrieved from http://www.bestevidence.org/

Bowman-Perrot, L. J., Greenwood, C. R., & Tapia, Y. (2007). The efficacy of CWPT used in secondary alternative school classrooms with small teacher/pupil ratios and students with emotional and behavioral disorders. *Education and Treatment of Children*, 20, 65-87.

Brantlinger, E., Jimenez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children*, *71,* 195-207.

Browder, D. M., Ahlgrim-Dezell, L., Spooner, F., Mims, P. J., & Baker, J. N. (2009). Using time delay to teach literacy to students with severe developmental disabilities. *Exceptional Children, 75,* 343-364.

Browder, D. M., Spooner, F., Ahlgrin-Delzell, L., Harris, A. A., & Wakeman, S. (2008). A meta-analysis on teaching mathematics to students with significant cognitive disabilities. *Exceptional Children, 74*, 407-432.

Browder, D. M., Wakeman, S. Y., Spooner, F., Ahlgrim-Dezell, L., & Algozzine, B. (2006). Research on reading instruction for individuals with significant cognitive disabilities. *Exceptional Children, 72,* 392-408.

Burks, M. (2004). Effects of classwide peer tutoring on the number of words spelled correctly by students with LD. *Intervention in School & Clinic, 39,* 301-304.

Burns, M. K. & Ysseldyke, J. E. (2009). Reported prevalence of evidence-based instructional practices in special education. *The Journal of Special Education, 43,* 3-11.

Byrd, D. E. (1990). Peer tutoring with the learning disabled: A critical review. *The Journal of Educational Research*, 84, 115-118.

Carnine, D. (1997). Bridging the research to practice Gap. *Exceptional Children, 63,* 513-521.

Chard, D. J., Ketterlin-Geller, L.R., Baker, S.K., Doabler, C., & Apichatabutra, C. (2009). Repeated reading intervention for students with learning disabilities: Status of the evidence. *Exceptional Children, 75,* 263-281.

Cook, B.G., Buysse, V., Klingner, J. Landrum, T. J., McWilliam, R., Tankersley, M., & Test, D. W. (2013) CEC's Standards for classifying the evidence base of practices in special education. Unpublished Manuscript

Cook, B.G. & Cook, S. C. (2013). Unraveling evidence based practices. *The Journal of Special Education*, 47, 71-82.

Cook, B.G. & Cook, S. C. (2011). *Thinking and communicating clearly about evidence-based practices in special education.* Division for Research White Paper.

Cook, B.G. & Schirmer, B. R. (2003). What is special about special education?: Overview and analysis. *The Journal of Special Education, 37,* 200-2005.

Cook, B. G. & Schirmer, B. R. (2006). An overview and analysis of the role of evidence-based practices in special education. In B.G Cook & B. R. Shirmer (Eds.), What is Special About Special Education?: Examining the Role of Evidence-Based Practices (pp. 175-185). Austin, TX: PRO-ED, Inc.

Cook, B.G. & Tankersley, M. (2007). A preliminary examination to identify the presence of quality indicators in experimental research in special education. In J. Crockett, M. M. Gerber, & T. J. Landrum (Eds.), *Achieving the radical reform of special education: Essays in honor of James M. Kauffman* (pp. 189-212). Mahwah, NJ: Lawrence Erlbaum.

Cook, B. G., Tankersley, M., Cook, L., & Landrum, T. J. (2008). Evidence-based practices in special education: Some practical considerations. *Intervention in School & Clinic, 44*, 69-75.

Cook, B.G., Tankersley, M., Landrum, T.J. (2009). Determining evidence based practices in special education. *Exceptional Children, 75,* 365-383.

Cook, B., Test, D., McWilliam, R., Buysse, V., Klingner, J., Landrum, T., & Tankersley, M. (2013). *Quality indicators and standards for determining evidence based practices in special education.* Unpublished Manuscript.

Cook, L., Cook, B. G., Landrum, T. J., Tankersley, M. (2008). Examining the role of group experimental research in establishing evidence-based practices. *Intervention in School and Clinic, 41,* 76-82.

Darling-Hammond, Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). National Staff Development Council and The School Redesign Network at Stanford University. *Professional Learning in the Learning Profession: A Status Report on Teacher Development in the United States and Abroad*. Retrieved from http://www.learningforwardpennsylvania.org/Professional%20Learning.pdf

Delquadri, J.C., Greenwood, C. R., Stretton, K., Hall, R. V. (1983). The peer tutoring spelling game: A classroom procedure for increasing opportunity to respond and spelling performance. *Education and Treatment of Children, 6,* 225-239.

Deschler , D. D. (2003). Intervention research and bridging the gap between research and practice. *Learning Disabilities: A Contemporary Journal, 1,* 1-7.

Detrich, R. (2008). Evidence-based, empirically supported, or best practice?: A guide for the scientist practitioner. *Effective Practices for Children with Autism: Educational and Behavioral Support Interventions that Work* (J.K. Luisell, D.C.

Russo, W.P Christian, S. M. Wilczynski, Eds). New York, NY: Oxford University Press.

DuPaul, G. J. & Henningson, P. N. (1993). Peer tutoring effects on the classroom performance of children with attention deficit hyperactivity disorder. *School Psychology Review, 22*, 134-143.

Dupaul ,G. J., Ervin, R. A., Hook, C. L. & McGoey, K. E. (1998). Peer tutoring for children with attention deficit hyperactivity disorder: Effects on classroom behavior and performance. *Journal of Applied Behavior Analysis, 31,* 579-592.

Flippin, M., Reszka, S. & Watson, L.R. (2010). Effectiveness of the picture exchange communication system (PECS) on communication and speech for children with autism spectrum disorders: A meta-analysis. *American Journal of Speech-Language Pathology, 19,* 178-195.

Fuchs, D. & Fuchs, L. S. (1995). What's 'special' about special education? *Phi Delta Kappan, 76,* 522-530.

Fuchs, D., Fuchs, L. S., Mathes, P. G., Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Journal, 34,* 174-206.

Fuchs, L. S. & Fuchs, D. (1995). Acquisition and transfer effects of classwide peer-assisted learning strategies in mathematics. *Schools Psychology Review, 24,* 604-21.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B .F., Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38,* 915-945.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71,* 149-164.

Greenwood, C. (1997). Classwide peer tutoring. *Behavior and Social Issues, 7,* 53-57.

Greenwood, C. R. & Abbott, M. (2001). The research to practice gap in special education. *Teacher Education and Special Education, 24,* 276-289.

Greenwood, C. R., Delquadri, J. C. & Carta, J. J. (1997). *Together we can!: Classwide peer tutoring to improve basic academic skill*. Longmont, CO: Sopris West Educational Services.

Greenwood, C. R., Delquadri, J. C., Stanley, S. p., Terry, B., & Hall, R. V. (1985). Assessment of ecobehavioral interaction in school settings. *Behavioral Assessment, 7*, 331-347.

Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M. E., Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice,15*, 198-205.

Guralnick, M. J. (1999). Second-generation research in the field of early intervention. In M. Guralnick (Ed.), *The effectiveness of Early Intervention (3-22).* Baltimore, MD: Paul Brookes.

Harper, G. F., Mallette, B., Maheady, L., Bentley, A. E., Moore, J. (1995). Retention and treatment failure in classwide peer tutoring: Implications for further research. *Journal of Behavioral Education, 5,* 399-414.

Harper, G.F., Mallete, B., Maheady, L., Parkes, V. & Moore, J. (1993). Retention and generalization of spelling words acquired using peer-mediated instructional procedures by children with mild handicapping conditions. *Journal of Behavioral Education, 3,* 25-38.

Harper, G. F., Mallete, B. & Moore, J. (1991). Peer-mediated instruction: Teaching spelling to primary school children with mild disabilities. *Reading, Writing, and Learning Disabilities, 7*, 137-151.

Higgins, J.P.T., Green, S. (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Holcombe, A., Wolery, M. & Gast. D. L. (1994). Comparative single-subject research description of designs and discussion of problems. *Topics in Early Childhood Special Education 14*, 119-145.

Horner, R. H., Carr, E .G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special

education. *Exceptional Children, 71*, 165-179.

Huberman, M. (1983). Recipes for busy kitchens: A situational analysis of routine knowledge used in school. *Knowledge, Creation, Diffusion, Utilization, 4,* 478-510.

Hughes, T. A. & Fredrick, L. D. (2006). Teaching vocabulary with students with learning disabilities using classwide peer tutoring and constant time delay. *Journal of Behavioral Education, 15,* 1-23.

Hume, K. & Odom, S. (2007) Effects of an individual work system on the independent functioning of students with autism. *Journal of Autism and Developmental Disorders 37*, 1166-1180.

Individuals With Disabilities Education Act(IDEA) of 2004, 20 U.S.C. 1415 *et seq.* (2004)

Jitrendra, A. K., Burgess. C., Gajria (2011). Cognitive strategy instruction for improving expository text comprehension of students with learning disabilities: The quality of evidence. *Exceptional Children, 77*, 135-159.

Jitendra, A. K., DuPaul, G. J., Someki, F. & Tresco, K. E. (2008). Enhancing academic achievement for children with attention-deficit hyperactivity disorder: Evidence from school-based intervention research. *Developmental Disabilities, 14,* 325-330.

Juni, P. Witschi, A. Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analyses. *The Journal of the American Medical Association, 282,* 1054-1060.

Kamps, D. M., Barbetta, P. M., Leonard, B. R., Delquadri, J. (1994). Classwide peer tutoring: An integration strategy to improve reading skills and promote peer interactions among students with autism and general education peers. *Journal of Applied Behavioral Analysis, 27,* 49-61.

Kennedy, C. H. (2005). *Single-Case Designs for Educational Research.* Boston: Pearson.

Kennedy, M. M. (1997). The connection between research and practice. *Educational Researcher, 26,* 4-12.

Kretlow, A.G. & Blatz, S. L. (2011). The ABCs of evidence-based practices for teachers. *TEACHING Exceptional Children, 43,* 8-19.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Landrum, T. J., Cook, B. G., Tankersley, M., Fitzgerald, S. (2002). Teacher perceptions of the trustworthiness, usability, and accessibility of information from different sources. *Remedial and Special Education, 23,* 42-48.

Lane, K. L., Kalberg, J. R., & Shepcaro, J. C. (2009). An examination of the evidence base for function-based interventions for students with emotional and/or behavioral disorders attending middle and high schools. *Exceptional Children, 75,* 321-342.

Lane, K. L., Harris, K. R., Graham, S., Weisenbach, J. L., Brindle, M., Morphy, P. (2008). The effects of self-regulated strategy development on the writing performance of second-grade students with behavioral and writing difficulties. *Journal of Special Education, 41,* 234-253.

Maheady, L. & Gard, J. (2010). Classwide peer tutoring: Practice, theory, research, and personal narrative. *Intervention in School and Clinic, 46*, 71-78.

Maheady, L. & Harper, G.F. (1987). A classwide peer tutoring program to improve the spelling test performance of low-income, third and fourth grade students. *Education and Treatment of Children, 10*, 120-133.

Maheady, L., Harper, G. F., & Mallette, B. (2001). Peer mediated instruction and interventions and students with mild disabilities. *Remedial and Special Educaiton, 22,* 4-14.

Maheady, L., Harper, G. F., Sacca, M. K. (1988a). Peer-mediated instruction: A promising approach to meeting the diverse needs of LD adolescents. Learning Disability Quarterly, 11, 108-113.

Maheady, L., Harper, G. F., Sacca, M. K. (1988b). The effects of classwide peer tutoring program on the social studies test performance of secondary, mildly handicapped students in resource room programs. *Journal of Research and Development in Education, 2,* 76-83.

Maheady, L.,Sacca, M. K. & Harper, G. F. (1988). Classwide peer tutoring with mildly handicapped high school students. *Exceptional Children, 55*, 52-59.

Maheady, L., Sacca, M. K., & Harper, G. F. (2001). Classwide student tutoring teams: The effects of peer-mediated instruction on the academic performance of secondary mainstreamed students. *The Journal of Special Education, 21,* 107-121.

Mastropieri, M. A., Scruggs, T. E., Norland, J. J., Berkely, S., McDuffie, K., Tornquist, E. H., Connors, N. (2006). Differentiated curriculum enhancement in inclusive middle school science: Effects on classroom and high stakes tests. *The Journal of Special Education, 40,* 130-137.

Mastropieri, M. A., Scruggs, T. E., Spencer, V., & Fontana, J. Promoting success in high school world history: Peer tutoring versus guided notes. *Learning Disabilities Research & Practice, 18,* 52-65.

Matheson, C. (1997). The effects of classwide peer tutoring on the academic achievement and classroom deportment of children with attention deficit hyperactivity disorder. (Doctoral dissertation) Retrieved from ProQuest Dissertations & Theses database. (9802703).

McDuffie, K. A. and Scruggs, T.E. (2008) The contributions of qualitative research to discussion of evidence-based practices in special education. *Intervention in School and Clinic, 44*, 91-97.

McMaster, K. L., Fuchs, D., & Fuchs, L. S. (2006). Research on peer-assisted learning strategies: The promise and limitations of peer mediated instruction. *Reading and Writing Quarterly, 22,* 5-25.

Montague, M. & Dietz, S. (2009). Evaluating the evidence base for cognitive strategy instruction and mathematical problem solving. *Exceptional Children, 75,* 285-302.

Mortweet, S. L., Utley, C. A., Walker, D., Dawson, H. L., Delquadri, J. C., Reddy, S. S., Greenwood, C.R., Hamilton, S., Ledford, D. (1999). Classwide peer tutoring: Teaching students with mild mental retardation in inclusive classrooms. *Exceptional Children, 65*, 524-536.

Moyer, A. & Finney, J. W. (2005). Rating methodological quality: toward improved assessment and investigation. *Accountability in Research, 12*, 299-313.

No Child Left Behind Act (NCLB) of 2001*,* P.L. No. 107-110, 115 Stat. 1425 *et seq*. (2001).

Odom, S. L. (2008). The tie that binds: Evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education, 29*, 53-61.

Odom, S. L., Brantlinger, E., Gersten, R. Horner, R. H., Thompson, B. & Harris, K.R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children 71,* 137-148.

Peters, M. T., & Heron, T. E. (1993). When the best is not good enough: An examination of best practice. *The Journal of Special Education, 26*, 371-375.

Promising Practice Network (2013). *Promising Practices Network.* RAND Corporation. Retrieved from http://www.promisingpractices.net/

Random.org. *True random number generator.* Retrieved from http://www.random.org/

Ryan, J. B., Reid, R. & Epstein, M. H. (2004). Peer-mediated intervention studies on academic achievement for students with EBD: A review. *Remedial and Special Education 25,* 330-341.

Sackett, D. L. (1997). Evidence based medicine. *Seminars in Perinatology, 21,* 3-5.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*, 71-72.

Sideridis, G. D., Utley, C., Greenwood, C. R., Delquadri, J., Dawson, H., Palmer, P., & Reddy, S. (1997). Classwide peer tutoring: effects on the spelling performance and social interactions of students with mild disabilities and their typical peers in an integrated instructional setting. *Journal of Behavioral Education, 7,* 435-462.

Slavin, R. E. (1989). PET and the pendulum: Faddism in education and how to stop it. *The Phi Delta Kappan, 70,* 752-758.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practices and research. *Educational Researcher, 31,* 15-21.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37,* 5-14.

Smith, G. J., Schmidt, M. M., Edelen-Smith, P. J. & Cook, B. G. (2013). Pasteur's quadrant as the bridge linking rigor with relevance. *Exceptional Children, 79,* 147-161.

Stenhoff, D. M. & Lignugardis/Kraft, B. (2007). A review of the effects of peer tutoring on students with mild disabilities in secondary setting. *Exceptional Children, 74*, 8-30.

Stevens, M. L. (1998). Effects of classwide peer tutoring on the classroom behavior and academic performance of students with ADHD. (doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (9902329).

Tankersley, M., Cook, B. G., & Cook, L. (2008). A preliminary examination to identify the presence of quality indicators in single subject research. *Education and Treatment of Children, 31,* 523-548.

Terry, B. (n.d.). ClassWide peer tutoring. *University of Kansas.* Retrieved from http://www.specialconnections.ku.edu/~kucrl/cgi-bin/drupal/?q=instruction/classwide_peer_tutoring

Test, D. W., Fowler, C. H., Brewer, D. M. & Wood, W. M. (2005). A content and methodological review of self-advocacy intervention studies. *Exceptional Children, 72,* 101-125.

Test, D. W., Fowler, C. H., Richter, S.M., White, J., Mazzotti, V., Walker, A.R., Kohler, P. & Kortering, L. (2009). Evidence-based practices in secondary transition. *Career Development and Transition for Exceptional Individuals, 32,* 115-128.

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71*, 181-194.

Tinkunoff, W. J. & Ward, B. A. (1983). Collaborative research on teaching. *The Elementary School Journal, 83,* 453-468.

Utley, C. A., Reddy, S. S., Delquadri, J. C., Greenwood, C. R., Mortweet, S. L., Bowman, V. (2001). ClassWide peer tutoring: An effective teaching procedure for facilitating the acquisition of health education and safety facts with students with developmental disabilities. *Education and Treatment of Children, 24,* 1-27.

Wanzek, J. & Vaughn, S. (2006). Bridging the research-to-practice gap: Maintaining the consistent implementation of research-based practices. In B.G Cook & B. R. Shirmer (Eds.), What is Special About Special Education?: Examining the Role of Evidence-Based Practices (pp. 165-174). Austin, TX: PRO-ED, Inc.

Weatherall, D. J. (1996). Foreword to the first edition by Professor Sir David Weatherall. In Greenhalgh (2010) (Author), *How to Read a Paper: The Basics of Evidence-Based Medicine* (pp. ix). West Sussex, UK: John Wiley & Sons.

What Works Clearinghouse (2011.). WWC intervention report: University of Chicago school mathematics project 6-12 curriculum. Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/WWC/interventionreport.aspx?sid=587

What Works Clearinghouse. (2011).*Procedures and standards handbook*

(version 3.0). Institute of Educational Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1

What Works Clearinghouse (2010). WWC intervention report: Classwide peer tutoring. Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_cwpt_091410.pdf

Viera, A. J. & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 35,* 360-363.

Yeaton, W. & Sechrest, L. (1981). Critical dimensions in the choice of maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156-167.