ABSTRACT

Title of Thesis:	A CASE STUDY OF THE CODING WORKBOOK FOR EVIDENCE-BASED INTERVENTIONS IN SCHOOL PSYCHOLOGY
	Lisa Henderson Sowers, Masters of Arts, 2005
Thesis directed by:	Professor William Strein Department of Counseling and Personnel Services

The *Procedural and Coding Manual* was developed by the Task Force in school psychology for the identification of psychological and educational intervention studies with empirical support (Kratochwill & Stoiber, 2002). The *Coding Workbook* (Shernoff & Kratochwill, 2004) was later designed to supplement the Manual with more explicit coding directions and illustrations for application of the coding criteria. The *Coding Workbook* has yet to be published or studied through formal research to determine its usefulness for the training of school psychologists.

The present study evaluated *Workbook* effectiveness by conducting the training and measuring pre-and post-training differences in inter-rater correspondence and coding time. In addition, reviewers evaluated the merits and challenges of the training process in detailed observation logs. Results of the study demonstrated that enhanced coding instructions and examples are needed to increase usefulness of the *Workbook* for training school psychologists.

A CASE STUDY OF THE CODING WORKBOOK FOR EVIDENCE-BASED INTERVENTIONS IN SCHOOL PSYCHOLOGY

by

Lisa Henderson Sowers

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
Of the requirements for the degree of
Master of Arts
2005

Advisory Committee:

Professor William Strein, Chair
Professor Cortland Lee
Professor Sylvia Rosenfield

ACKNOWLEDGEMENTS

I wish to thank my colleagues for their in this participation in this project. I owe a debt of gratitude to Sara Signor, who made a significant contribution to this case study. I would also like to thank the school professionals at Thomas Johnson Elementary School. Through their support, I was able to excel as a member of the school team while also working on this research project.

I also want to thank each member of my committee. Drs. Sylvia Rosenfield and Cortland Lee expressed unwavering enthusiasm for this project and the importance of this contribution to the EBI movement in school psychology. My advisor, Dr. William Strein also expressed appreciation for this project as well as steadfast support for my research.

Most of all, I would like to thank my family. My academic success has been a reflection of their continued support. Special thanks to my husband, Kevin, whose support and patience made it possible for me to persevere. Finally, I dedicate this accomplishment to my mother, Patricia, who remains a source of strength in my life.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Background of the Study

Escalating health care costs in the 1980's urged managed care companies to implement cost containment measures (Beulter, 1998). In the following decade, establishing a national healthcare system in the U.S. moved to the forefront of the country's agenda. These events initiated an investigation of interventions used by the medical profession. Insurance companies conducting this assessment observed that a wide variety of medical procedures were used for treatment of specific health problems. Based on this idea, it was concluded that physicians were guided more by clinical judgment than they were by scientific information (Reed, McLaughlin & Newman, 2002). In response to these criticisms, the medical profession was the first in the U.S. to form a task force for the purpose of promoting evidence-based practice guidelines (Stoiber & Wass, 2002). Initiatives of the evidence-based intervention (EBI) movement in health care were directed at reducing variation in medical procedures, developing standards for application of methods and ultimately improving the efficiency of health care (Levant, 2004).

Managed care companies' interests in cost reduction also had a significant influence on the mental health profession. Within this era of standardized health care, the managed care system adopted guidelines that emphasized psychological treatment by less experienced service providers and reduced service fees (Beulter, 1998). These measures posed substantial threats to the quality of psychotherapy.

Other pressures from political and social forces mounted for an identification of effective psychotherapies approaches. At that time, over 400 interventions were employed in practice (Beutler, 1998). In this environment, the clinical psychology field faced the risk that the identification of reimbursable interventions would be determined by outside agencies. As a way of reclaiming control of these efforts, the American Psychological Association (APA) appointed the Task Force on Psychological Intervention Guidelines in 1985 (Chambless and Ollendick, 2001). The APA Division 12 Task Force was charged with the mission to disseminate information about effective treatments to clinical psychologists, third-party payers and the public (Levant, 2004). This group, utilizing the expertise of major figures in the field, also set out to ensure that clinical psychologists had the expertise to apply EBIs to everyday practice. Initiatives were established to promote integration of EBIs into graduate training programs and continuing professional development seminars (APA, Division 12, 1993).

In 1995, the Division 12 Task Force published the *Template for Developing Guidelines: Interventions for Mental Disorder and Psychosocial Aspects of Disorders* (APA Division 12 Task Force, 1995 as cited in Hughes, 2000). The report distinguished two categories of coding criteria and supplemented this information with 26 examples of programs meeting these criteria. Standards originally established by the Federal Drug Administration were adapted as criteria used to judge intervention efficacy (Beutler, 1998). *Well Established* studies are supported by evidence of superiority for an experimental group over a control sample in at least two randomized trials. Efficacy demonstrated in one randomized trial is utilized as evidence for the second category, *Probably Efficacious* interventions (Chambless & Ollendick, 2001). The list of approved treatments, predominated by behavioral and cognitive-behavioral therapies, promoted a manualized approach to clinical practice.

The efficacy of psychotherapy has always been a subject of debate in the field of clinical psychology. Deliberation among psychologists began with Hans Eysenck's inquiry of this topic in 1952 (Stoiber & Waas, 2002). Consequently, the question of what practice standards should be adopted in psychotherapy has remained controversial. The publication of the Division 12 Task Force Report sparked renewed debate of this topic among researchers and practitioners. Critics who examined the use of standardized practice guidelines found that implementation of this approach was either premature (Garfield, 1998) or lacked a proper emphasis on many critical factors contributing to intervention effectiveness (Ahn & Wampold, 2001; Beutler, 1998; Henry 1998; Norcross, 2001). The issues raised by these critics are summarized as follows: a) the review comprised a small and unrepresentative number of studies; b) the focus of the criteria did not represent approaches in the research literature or those commonly used in practice; c) the standard for empirical support based on randomized clinical trials did not account for therapist effects or client characteristics that contribute to positive therapy outcomes; and d) the Division 12 recommendations were likely to result in a limitation to approved interventions for reimbursement by managed care companies.

Soon after the establishment of an EBI movement in clinical psychology, a similar movement began to gain momentum in the field of school psychology. Rather than operating from pressures imposed by the managed care industry, this group drew their motivations from an interest in advancing the quality of children's psychological services and a desire to meet federal standards for evidence-based practices (Stoiber & Kratochwill, 2000). However, the field of school psychology shared other objectives with the medical and clinical psychology professions. With the purpose of disseminating information about effective interventions for children and youth, the Task Force on Evidence-Based Interventions in School Psychology (hereafter referred to as the Task Force) was founded in association with Division 16 of the APA and the Society for the Study of School Psychology (Gutkin, 2002). The group has also received the endorsement of the National Association of School Psychologists (NASP).

The initial mission of the Task Force focused on determining effective interventions for addressing the specific academic and behavioral problems of children in schools (Stiober & Waas, 2002). As schools are the largest provider of mental health care services to school-age children, this agenda remains critical to the profession of school psychology and education (Kratochwill & Shernoff, 2004). Increased incidences of problems encountered by school-age children emphasize the need for implementing quality standards for school-based service delivery.

Plans initially developed by the Task Force included the formation of an EBI knowledge database to disseminate information about EBIs in graduate training and clinical practice settings. Regenerating a beneficial exchange between the research and clinical communities was another key objective of the group (Kratochwill & Shernoff, 2004). As the first step toward accomplishing these goals, the group developed the *Procedural and Coding Manual for Review of Evidence-Based Interventions* (henceforth called the Procedural and Coding Manual or the Manual). The Manual defines criteria for identification, review and coding of psychological and educational intervention studies. Research domains, which comprise a focus on social/emotional, academic, and health care interventions, reflect the wide range of problems encountered by school-age children and their families (Stoiber & Waas, 2002).

The Task Force has encountered numerous challenges in its development of the Procedural and Coding Manual. Reviews of the document yielded both praise and criticism from experts in the field. A number of strong points associated with the Manual were reported by Levin (2002). These notable features include: a) the potential for application by the research and practice communities as a comprehensive resource for empirical research validation, b) consideration of clinical and educational significance of a particular study, and c) provision of information about key outcomes associated with a particular intervention technique. Reviewers have praised the Task Force for developing the most comprehensive tool for evaluating school-based research to date (Gutkin, 2002; Levin, 2002).

To a greater degree, the reviews of the Procedural and Coding Manual emphasized challenges experienced in its practical application. A number of factors have been identified as leading to decrements in the validity of coding results. The evaluator's knowledge of the research domain and competency in research methodology are some of these challenges (Christenson, et al. 2002). Another difficulty concerns the level to which inferential judgment is necessitated by missing or vaguely stated information (Levin, 2002). Inferences about indistinct information were observed to result in low rates of inter-rater agreement.

Coding studies with the Manual has also required a considerable investment of time and energy. Prevatt and Kelly (2004) reported that their review of 18 studies took 80-90 hours. There is a significant likelihood that a graduate student or a school psychologist with less technical knowledge would need even more time to conduct a similar review. Adequate training would also be required to facilitate this task. This possibility raises concerns about the feasibility of the current coding scheme when applied to graduate level training programs or other practical settings.

Other issues have emerged with regard to developing standards for evaluating the merits of school-based research. The most salient concern centers on the defining "the level of controlled research that constitutes evidence" (Nelson & Epstein, 2002). Criteria based on randomized design established as the "gold standard" in other branches of psychology were considered by the Task Force during their initial conceptualization of the Manual. Although this standard was suitable for laboratory research, it was deemed as inappropriate for research commonly conducted in schools. To develop Manual criteria, the Task Force focused on creating a structure that reflected a broad view of evidence-based practice. As such, the Manual coding scheme addresses contextual factors and integrates principles of sound science. Coding criteria have been included for both randomized and nonrandomized design studies. Through these efforts, the Task Force intended to create an instrument for evaluating EBIs that could be useful to both research and practice (Stoiber, 2002). Despite these objectives, the decision to include criteria for experimental and non-experimental research designs has received criticism from the field.

The Task Force addressed issues raised by reviewers with numerous revisions to the Manual. One of these changes included the integration of prevention programs into the five general coding domains. Another comprised the addition of "confidence ratings" that denote the confidence level of the reviewer. The Task Force has also given additional consideration to the standards that should constitute an empirical basis for evaluating studies in school psychology. Although the Task Forces promotes randomized design as the "gold standard" for empirical research, the organization still included criteria for a variety of research designs in the Manual. Stoiber (2002) asserts that these criteria accommodate the realities of conducting experimental research in schools. Furthermore each approach is founded on sound scientific principles. According to the authors (The Task Force for Evidence-Based Interventions, 2003), other aspects of the Manual, such as the organization of research domains, are still subject to change.

Recently a number of other organizations have instituted initiatives to identify evidence-based interventions for school-age children.   One organization, the What Works Clearinghouse, has secured support from the U.S. Department of Education to establish large-scale efforts for this purpose. The mission of the What Works Clearinghouse (WWC) is to identify effective intervention, prevention and educational programs that are of particular relevance to the education community (What Works Clearing House, 2004). The substantial financial and professional resources utilized by this organization have the potential to facilitate an efficient evaluation of intervention studies that has been unmatched in the field.

Current issues surrounding the identification of effective interventions in school psychology have encouraged the Task Force to revisit their mission. As a part of this effort, the group has formulated new directions for the scope of their work. These examinations have resulted in the mission to improve the quality of research training and to promote the implementation of EBIs in practice. This goal involves revising the Manual and creating other tools that are useful for practice and training environments. A major challenge of this objective was to develop tools that were effective in preparing school psychologists for evidence-based practice.

The Coding Workbook for Evidence-Based Interventions was developed to achieve this objective. It was designed to supplement the Manual with more explicit coding directions and to provide illustrations of the coding procedure (Kratochwill 2002). The Workbook, similar to the Coding Manual, was initially developed for use by the Task Force which is largely comprised of faculty members at research institutions. More recently, the Workbook has been recognized as potentially useful for graduate training in EBIs.

The Workbook is also part of an initiative to provide expanded training for all potential consumers of the Manual. As part of this effort, the Task Force aims to identify factors that contribute to inter-rater correspondence and use this information to formulate better training procedures (Kratochwill & Shernoff, 2004). The Coding Workbook (Shernoff & Kratochwill, 2004) comprises a series of exercises that serve as self-instructional training for coding research using criteria developed by the Task Force. Coding illustrations for various dimensions and criteria in the Procedural and Coding Manual are outlined. The Workbook has yet to be published by the Task Force or be examined in formal research.

The Future directions also include building a more collaborative relationship with practitioners. Kratochwill (2004) believes that this alliance will result in practitioner's increased input into the EBI research agenda. Forming this alliance is also critical to ensuring practitioner's acceptance of EBIs as relevant to their practice and beneficial to their ultimate clients. Despite the considerable work accomplished by the Task Force, the group is still striving to achieve its long-term goals. Kratochwill (2002) states that while encouragement is more conducive to the ongoing efforts of the group, "we remain open to criticism, suggestions, and feedback".

Purpose of the Study

The purpose of the present study was to evaluate the usefulness of *the Coding Workbook for Evidence-Based Interventions* in School Psychology. The case study examined the extent to which the Workbook was effective in increasing reviewers' proficiency in coding Manual criteria. Quantitative data was generated by conducting the training and measuring differences in pre-and post-training inter-rater correspondence. Reviewer agreement to codes determined as "accurate" by a consensus of researchers was also examined. In addition, strengths and limitations of the Workbook exercises were documented in detailed observation logs taken throughout all phases of the project. The descriptive analysis generated by the study includes suggestions for improvements to the Workbook. These recommendations may facilitate further development of the Workbook as an effective training tool for school psychologists. The final report may also contribute to the evidence-based intervention literature.

Significance of the Study

As part of the initiative to improve the quality of service delivery in school psychology, the Task Force has focused on implementing EBIs in schools and other practical settings. This activity will provide critical information about the practice utility of EBIs. Kratochwill and Shernoff (2003) suggest that interventions identified as *efficacious* in research environments must also be proven as *effective* in practice. Collaboration with practitioners is critical to achieving this objective.   It is also essential to research directed at the transportability of efficacious approaches. Studies in this area examine "How the intervention works in the real world and who *can* and who *will* conduct the intervention, under what conditions and to what effect" (Schoenwald & Hoagwood, 2001 as cited in Kratochwill and Shernoff, 2003).

The challenges to this mission that still lie ahead are evidenced by the ongoing gap between research and practice. Beliefs held by school professionals about the potential effectiveness of EBIs in a particular school as well as their allegiance to particular intervention approaches have been identified as contributing factors to this failed alliance (The Evidence-Based Intervention Work Group, 2005). Administrative and practical barriers that exist in school settings may also interfere with daily integration of EBIs. Kratochwill and Shernoff (2003) have observed that the additional time and resources necessary to access information about evidence-based methods serve as obstacles to their implementation. The technical nature of traditionally-drafted research studies may also account for this gap. In addition to limited access, school professionals may lack the knowledge of research methodology necessary to interpret these research studies (Kratochwill & Shernoff, 2004).

The Task Force plans to facilitate the implementation of EBIs in practice by providing the field with the tools and training necessary to promote school psychologists' competence in scientific research. An essential part of this effort focuses on increasing graduate training in interventions with proven effectiveness (Kratochwill and Shernoff, 2003). The Workbook, which supplements the

Manual by illustrating practical application of the coding criteria, can be instrumental in this initiative. The Workbook contains coding exercises and predetermined answers that pertain to a variety of school-based studies. As such, it may be useful in the instruction of criteria that defines “evidence-based” and methods for evaluating interventions used in practice.

Research that has been conducted to measure a practical application of the criteria has focus on the Procedural and Coding Manual. Some of these studies indicated that coding with the Manual requires a substantial background in research methodology. Of the four reviews identified for this project, two were conducted by members of the Task Force (Lewis-Snyder, Stoiber, & Kratochwill, 2002; Shernoff, Kratochwill & Stoiber, 2002). Because the authors were also members of the Manual subcommittee, they had a well developed understanding of its underlying theoretical and methodological foundations. The third review, authored by researchers independent of the Task Force, noted many issues with regard to the criteria. These issues included a) the discernment of coding criteria, particularly in areas of design characteristics, b) interpretation of Statistical Treatment criteria, and c) identifying intervention components (Prevatt & Kelly, 2004). Additionally, Master’s level graduate students experienced difficulties with interpreting the Procedural and Coding Manual that prevented their further participation in the research project. In their comments, Prevatt and Kelly (2004) state that “there is legitimate concern that researchers (and particularly graduate students) will not have sufficient technical expertise to reliably code studies and that level of expertise will make many trainers reluctant to include the Manual as part of their standard school psychology training". The conclusions drawn by Prevatt and Kelly (2004) based on their research of the Manual has significant implications for the usefulness of the Coding Workbook.

A number of other reviewers have identified challenges associated with the Procedural and Coding Manual. These issues include the following: a) establishing adequate rates of inter-rater agreement (Levin, 2002; Prevatt & Kelly, 2004; Christenson, et al., 2002), b) conducting a representative literature review (Nelson & Epstein, 2002), c) understanding the level of positive coding results that constitutes evidence (Nelson & Epstein, 2002), and d) conducting an appropriate analysis of contextual factors (Christenson et al., 2002). The problems that have been encountered with application of the Manual may limit its broad use in practical settings. Because the same criteria are employed in the Coding Workbook, these issues are also relevant to its usefulness as a training instrument for school psychologists. These issues must be addressed in order for the Task Force to implement the Manual and Workbook.

The present study may be of particular relevance to the Task Force because it served to illustrate challenges associated with use of the Coding Workbook. It provides important information for development of the Workbook by demonstrating its usefulness in training reviewers with different levels of research experience. The potential for Specialist level graduate students trained in the scientist-practitioner orientation to conduct Workbook training was examined in this study. Reviewers' ability to code studies with the Manual following use of the Workbook was also investigated. Prevatt and Kelly (2004) state that the considerable time investment and technical expertise required by the Coding Manual may result in its subsequent exclusion from many graduate school programs. Information generated from this study attempted to address these issues and made recommendations for its effective use in graduate student training programs. Additionally, the study demonstrated the level of knowledge in research methodology obtained by the graduate students, which may provide a basis for designing training standards tailored to similar consumers of EBI research.

The study is also intended to aid Task Force efforts focused on structuring instruments that are useful to trainers and research psychologists. Christenson et al. (2002) found that substantial background in research methodology was required to conduct valid interpretations of Manual criteria. The usefulness of Workbook training when conducted by a faculty member with considerable research experience was investigated in this study. Practical application of the Manual by reviewers in pre-and post-training conditions was also demonstrated. Results generated by the study may inform the Task Force about application of the Workbook by reviewers at various levels of experience.

In summary, my research may contribute to the ultimate goal of Task Force which involves promoting the application of a scientist-practitioner orientation in practical settings. As part of this initiative, the group recommends that practitioners use the Procedural and Coding Manual to guide their everyday practice. However, the technical nature of the current instructions may present obstacles to achieving this goal. My research focuses on a Workbook designed to promote school psychologist's knowledge of the Manual criteria. The case study may provide the Task Force with information for enhancing the usefulness of this instrument.

Statement of the Problem

Implementation of evidence-based interventions by school psychologists is necessary to ensure the ongoing improvement of education in schools. To meet this objective, the Task Force has embraced the EBI movement and currently focuses its efforts on facilitating the adoption of EBIs in schools and other prevention settings (The Evidence-Based Intervention Work Group, 2005). Providing school psychologists with the background necessary to interpret and evaluate scientific research will be essential to promoting their adoption of new intervention methods. Efforts to develop training and instruments that broaden the applicability of the Coding Manual have been begun by the Task Force. *The Coding Workbook for Evidence-Based Interventions* in School Psychology has been developed as part of this initiative.

The Task Force is still in the process of finalizing the Workbook. The challenges and issues associated with this instrument must be documented before future revisions can be designed and employed. Therefore, research is needed to illustrate the feasibility of the training and usefulness of the Workbook outcomes.

This study attempted to evaluate Workbook usefulness by conducting the self-instructional training and measuring differences in pre-and post training coding performance. It focused on identifying the strengths and limitations of the

Workbook exercises based on quantitative and qualitative data collected prior to and following training. It initially investigates reviewer coding accuracy for Workbook exercises and commentary about conducting the training module. Next, application of the Coding Manual prior to and following Workbook training is examined. These results are organized by the following research questions: a) is there an increase in inter-rater correspondence following use of the Coding Workbook?, b) is there an increase in consensus code agreement following use of the Workbook training?, c) is there a decrease in the time required to code studies following Workbook instruction? Finally, the study summarizes reviewer commentary about the Coding Manual. It examines observation logs recorded prior to and following Workbook training that provide important information about utility of the Workbook.

Definition of Variables

Evidence-based Interventions

Evidence-based interventions are grounded on prior research findings as effective when performed with a particular population or clinical setting (Kratochwill & Shernoff, 2002). The objective of the Task Force is to promote evidence-based practice in schools by providing guidelines for field-based research. Efforts to identify EBIs for school-age children are currently being conducted by other groups such as the What Works Clearinghouse (www.w-w-c.org). In this study, the term evidence-based intervention will be used to refer to empirically-based methods in school psychology and its variants (empirically-validated treatments, empirically supported therapies, scientific therapies) that appear in the educational, medical and broader mental health disciplines.

Task Force

The Task Force on Evidence-Based Interventions in School Psychology was established to create a knowledge based on effective intervention and prevention programs in the field of school psychology (Kratochwill & Stoiber, 2002; Stoiber & Waas, 2002). It employs the diverse talents of professionals from education, psychology and other related fields to identify, review and code studies of behavioral and academic interventions for children. As part of their mission, the Task Force will disseminate EBI information to the school psychology community and advance the use of scientific research in the field (Kratochwill & Shernoff, 2004). The work of the Task Force has been supported by APA Division 16, the Society for the Study of School psychology and endorsed by the National Association of School Psychologists (Kratochwill & Stoiber, 2002; Prevatt & Kelly, 2004).

Procedural and Coding Manual

The Procedural and Coding Manual for the Review of Evidence-Based Interventions describes coding criteria developed by the Task Force for evaluating the empirical support of intervention and prevention studies (Kratochwill & Stoiber, 2002). The Manual is divided into four sections corresponding to different research designs: a) group-based design, b) single participant design, c) qualitative research methodology, and d) confirmatory program evaluation. Studies are examined based on general characteristics, key evidence components, and other descriptive or supplemental features (Stoiber & Waas, 2002). Thus, the Manual constitutes a focus on the empirical/theoretical basis and statistical properties of interventions. Consideration of important external and internal factors is also integrated into the coding scheme.

Coding Workbook

The Coding Workbook (Shernoff & Kratochwill, 2004) was designed to supplement the Manual with actual illustrations of the coding criteria. It provides a self-instructional training to assist individuals in learning about the criteria and coding research investigations. Exercises in the Workbook comprise study excerpts, coding for a variety of study dimensions and pre-determined coding responses. Single-Participant and Group-Based Design studies are investigated based on General Characteristics, Key Features and Other Descriptive criteria.

Educational/Clinical Significance

Educational/clinical significance is the degree to which the effects of intervention are meaningful in the context of the implementation setting (Chambless & Hollon, 1998). Furthermore, interventions must be useful to the practitioner and beneficial to the particular presenting problem of the client. The Procedural and Coding Manual criteria demonstrate clinically significant outcomes produced by an intervention, such as reduction of problem behavior and information about meaningful changes in instructional outcomes (Lewis-Synder, Stoiber, & Kratochwill, 2002). Both the Key Evidence and Other Descriptive Information sections of the Manual include consideration of internal and external validity indicators such as participant characteristics and the site of implementation that are essential to assessing educational/clinical significance. Furthermore, it provides consumers with information necessary to evaluate the appropriateness of the intervention to their specific needs (Stoiber & Waas, 2002).

The Promoting Alternative THinking Skills Program (PATHS)

The Promoting Alternative THinking Skills Program (PATHS) is an intervention program aimed at enhancing children's social and emotional functioning. It comprises a series of structured classroom activities that focus on recognizing emotions, developing self-management strategies and practicing these strategies in challenging interpersonal situations (Seifer et al., 2004). The program is typically implemented on a school-wide level by teachers with the support of an on-site coordinator and other research staff. The PATHS curriculum has proven utility for addressing children's social/emotional concerns, particularly students from diverse cultural backgrounds (Greenberg, Kam & Walls, 2003; Greenberg, Kusche, Cook & Quamma, 1995; Seifer et al., 2004).

PATHS research was selected for this project because studies of this program are highly applicable to criteria in the Coding Manual. Stoiber (2002) asserted that the Manual has been structured to accommodate research that is commonly conducted in schools. PATHS research is typically school-based and employs group-based designs common for research performed in these settings. The studies utilize random and nonrandom group assignment methods that are relevant to review with Manual criteria.

CHAPTER 2

REVIEW OF THE LITERATURE

This chapter will review the history of the evidence-based intervention movement. First, the movement that began with formation of a task force in the medical and clinical psychology professions is presented. Next, a similar trend in the field of school psychology is introduced with a discussion of its merits and critical issues. The evolving mission of the Task Force directed at implementing EBIs in practice environments and the potential outcomes of these efforts are considered.

The History of Evidence-Based Interventions

Responding to pressures from managed care companies to contain treatment costs, medical professionals were the first to establish guidelines for treatment according to diagnosis (Stoiber & Waas, 2002). This movement that originated in the United Kingdom was quickly adopted by the U. S. medical profession. Growing concerns about health issues and other problems confronted by society promoted interest by several groups to identify effective treatments. For example, The Center for Disease Control (http://www.cdc.gov/nccdphp/dash/rtc/index.htm) focused efforts on reduction of HIV, STD and pregnancy while crime and delinquency prevention was investigated by Loeber, at el., (1999). The movement towards establishing the empirical basis of medical interventions and pressures from managed care companies to contain rising costs of therapy prompted the need for service delivery guidelines in clinical psychology. As concerns about mental health costs grew, psychologists were faced with the potential that outside agencies, such as managed care companies or the federal government, would establish standards for reimbursable treatment.

In 1995, Division 12 of the APA appointed the Task Force on Psychological Intervention Guidelines (Chambless and Ollendick, 2001). It combined the expertise of clinical psychologists from academic, medical and clinical sectors to evaluate the empirical basis of treatments commonly used and considered efficacious by the field. The evidence-based movement in clinical psychology mirrored premises that originated in the medical field: a) training in current empirically validated treatments will enhance client care, b) the busy schedules of clinicians limit their ability to remain abreast of the most up-to-date treatments c) clinician's failure to supplement their empirical knowledge leads to decrements in practice, d) summaries of empirically-based treatments supplied by expert reviewers, complete with instructions of how interventions can be easily applied to daily practice, would be beneficial to clinicians (Chambless & Ollendick, 2001).

One goal of the Division 12 Task Force was to ensure that therapies approved for reimbursement by managed care companies included scientifically proven methods (APA Division 12 Task Force, 1993; Chambless, 1998). Formerly, interventions selected on the basis of past experience or training exposure were commonly submitted for third-party billing (Kendall, 1998). Related concerns arose with regard to training standards in clinical psychology. Based on their knowledge of APA-approved programs, the Division 12 Task

Force suspected that graduate level training in clinical psychology lacked an adequate focus on EBIs. If graduate students failed to gain this necessary exposure, they would be ill-prepared for the application of empirically-based methods to daily practice. Furthermore, the Division 12 Task Force feared that the clinician's practice would be confined to the treatments they learned during graduate training that comprised a small percentage of scientifically supported treatments. Therefore, conducting an evaluation of EBIs integrated as part of APA graduate programs was adopted as a key initiative of the Task Force (APA Division 12 Task Force, 1993).

In a survey of clinical graduate school directors, the Division 12 Task Force determined that 22 percent of the APA-approved programs provided instruction in less than 25 percent of the EBIs. Practicum students in 44 percent of APA approved clinical psychology programs received training in treatments on a preliminary list compiled by members (APA Division 12 Task Force, 1993). Although their research indicated that graduate programs provided some training in EBIs, the Task Force intended to increase integration of effective treatments in both graduate coursework and supervised clinical experiences of APA-approved programs. To meet this objective, the organization identified training in EBIs as an important criterion for accreditation of doctoral programs in clinical psychology. The guidelines stipulated that programs recruit faculty with expertise in empirically-based methods and extend current training to include the use of EBI treatment manuals. As part of the report published in 1993, the Division 12 Task Force also emphasized that continuing education programs for practicing clinical psychologists enforce current guidelines for training in empirically validated treatments.

In its 1993 report, the Task Force on Psychological Intervention Guidelines published criteria for the selection of EBIs and provided 19 examples of *well established* programs and 7 examples of *probably efficacious* programs. The purpose of the review was to provide an initial list of the scientifically-based treatments currently employed in the practice of clinical psychology. The initial criteria for evaluation of interventions were judged by the organization to be "somewhat arbitrary". Furthermore, the report indicated that other criteria absent from the list may be of equal importance.

Although hastily developed, the list was drafted for the purpose of surveying graduate school directors about EBI implementation. Studies reviewed by the Division 12 Task Force fell into two categories: *Well Established* and *Probably Efficacious Treatments* (APA Division 12 Task Force, 1993)**.** To be considered *Well Established,* an intervention must have comprised at least two randomized trials each with an adequate sample demonstrating superiority in evidence for the experimental group in comparison to the control group. Interventions that fall in the *Probably Efficacious* category are supported by at least one randomized controlled trial demonstrating superiority over control conditions or another valid treatment. Therefore, these programs are assessed at a level just below the criteria for *Well Established* treatments. By 1998, the list had grown to include 71 interventions (Chambless & Olendick, 2001). Division 12 has remained responsible for the evaluation of psychological intervention effectiveness. *The Procedural and Coding Manual for Identification of Beneficial Treatments* was produced in 2000 as a result of their efforts (Weisz & Hawley, 2000 as cited in Prevatt & Kelly, 2004). More recent searches of the literature and inspection of the APA Division 12 website indicate that the list of EBIs remains consistent at 71 programs.

To expand identification process beyond its focus on adults, the Division 12 assigned a second task force to investigate EBIs and prevention programs for children (Chambless & Ollendick, 2001). In 1998, the Task Force on Effective Psychosocial Interventions: A lifespan Perspective published research identifying a number of effective treatments (Weitz & Hawley, 1998). The review of over 300 therapies for children at various ages reported mean effects comparable to that found for adult psychotherapy as well as durable treatment outcomes. However, when the team specifically reviewed interventions used in clinical practice, dissimilar results were found. A second evaluation of nine studies found in the literature resulted in negligible effect sizes. The research, which called into question the effectiveness of conventional therapies used in clinical settings, underscored the urgent need for identification of beneficial interventions by the Division 12 Task Force. In 1998, the work of a third task force resulted in a published book entitled, *A Guide to Treatments That Work* (Nathan & Gorman, 1998 as cited in Chambliss & Ollendick, 2001). The book comprises reviews of psychotherapy and pharmacology studies by experts from these fields. Additionally, examinations of interventions in the area of adult, child, marital and family therapy have been conducted by groups independent of the APA (e.g., Kendall & Chambless, 1998). Psychologists in the United Kingdom and Canada have also contributed substantially to the identification of EBIs.

The EBI movement, along with pressures imposed by funding agencies and insurance companies, resulted in corresponding changes in psychology research. To meet growing demands of demonstrating study utility, researchers adopted methodologies as part of a new paradigm based on diagnosis-based clinical trials, standardized therapies and empirical validation. The acceptance of a scientific model has provided a number of advantages: a) an ease of study replication with greater reliability, b) evidence of the efficacy of a treatment, c) identification of the superiority of one modality over another for treatment of a particular disorder (Hibbs, 2001).

While many clinical psychologists supported the EBI guidelines and new approaches for psychology research, others have criticized these efforts (Beutler, 1998; Henry, 1998; Hibbs, 2001; Norcross, 2001). Chambliss & Olendick, (2001) propose that some professionals favored qualitative over quantitative research as a paradigm for psychotherapy research. In addition, many clinical psychologists failed to recognize individual techniques as responsible for intervention outcomes and therefore questioned the necessity of the identification process. In contrast, specific principles with proven effectiveness, common to a variety of approaches, were considered to be responsible for positive psychotherapy outcomes (Garfield, 1998). Still others characterized the review as limited in scope, comprising a small and unrepresentative segment of the research literature. This sample was thought to misrepresent the broad range of findings available in the literature (Beutler, 1998). Despite its claims of furthering scientific empiricism in clinical psychological research, endeavors by the Division 12 Task Force has been described as antiscientific and contradictory this objective (Henry, 1998).

The movement towards intervention guidelines was viewed by clinical psychologists as methods that underestimated the value of mental health services and limited the creative flexibility of practitioners. This criticism of the EBI approach largely related to its emphasis on standardized intervention manuals. The use of prescribed manuals was characterized as reducing psychotherapy services to an “automated” application of procedures by technicians (Hibbs, 2001). Thus, reliance on these manuals was considered as potentially leading to decrements in service quality.

Relatedly, manualized therapy was reported to ignore etiological factors and the contribution of individual personality factors to behavior. Norcross (2001) stated that differences in psychotherapy outcomes vary according to client characteristics. The impact of factors such as cross-diagnostic client characteristics, treatment goals, coping styles, stages of change, personality dimensions, and reactance level are disregarded in the suggestion that therapists dutifully follow treatment manual guidelines.

Norcross (2001) also argues that both clinical experience and research support the therapeutic relationship as the most influential factor in psychotherapy research. *Well Established* interventions in the EBI approach are presented to the psychology community as “what works” in the remediation of a particular concern. However, research by Lambert and Barley (2001 as cited in Levant, 2004) shows that the specific therapy technique or type only accounts for 15% of variance in therapy outcomes. Conversely, the relationship between therapist and client, along with other contextual factors, accounted for at least 30% of variance in treatment outcomes. Other research has revealed significant differences in the level of treatment effectiveness achieved by individual clinicians using the same treatment manual (Ahn, & Wampold, 2001; Crits-Christoph & Mintz, 1991). These variances may still exist when adherence to treatment manuals is confirmed (Henry, Strupp, Butler, Schacht, & Binder, 1993 as cited in Henry, 1998).

Issues were also raised with regard to the over-reliance on clinical trials in the evaluation of psychotherapies (Prevatt & Kelly, 2004). When research fails to address problems common to clinical settings, there is limited ability to generalize study outcomes to practical settings. The presence of comorbid conditions was identified as a salient concern in this regard. Laboratory settings provide opportunities to control for comorbid conditions that are not possible in clinical practice (Levant, 2004). The psychology practice in schools includes other factors that confound clinical research. Noncompliant parents and high absenteeism are identified as just some of these factors (Hibbs, 2001). Concerns have also been identified with regard to the failure of laboratory research to investigate long-term outcomes of problem behaviors (Prevatt & Kelly, 2004).

Additional issues have been raised about interventions labeled by the Division 12 Task Force as experimental or *Probably Efficacious.* Many of the methods commonly used in clinical settings (i.e. brief dynamic therapy) are characterized as failing to meet the rigorous standards of *Well Established* treatments. Opponents to the categorization system believe that treatments identified as *Well Established* are no more effective than those receiving the label of *Probably Efficacious* or “unproven”. There were also concerns that managed care companies will view *Probably Efficacious* treatments as experimental causing subsequent marginalization of these methods by clinicians and third-party payers (Henry, 1998). Since many of the approaches were required for internships and commonly used in practice, clinicians considered this issue of supreme importance.

Other concerns relate to the formation of an alliance between the Task Force and the managed care industry. Some professionals in the field viewed the identification of EBIs by the Division 12 Task Force as simultaneously strengthening their political connections with insurance companies (Henry, 1998). Exclusive reimbursement of EBIs by health care companies would also function to determine which treatments would be most commonly employed in practice. This potential transfer of power incensed psychologists who already considered the system to be intrusive, restrictive and insensitive to client’s best interests.

Many of these criticisms have been addressed by supporters of the empirically-based intervention movement. Chambless and Ollendick (2002) posit that much of the EBI controversy stems from the long standing existence of opposing views in the field. In light of issues raised by the opposition, the authors argue in support of manualized treatment and substantiate this argument with evidence. In two studies of interventions for anxiety disorder, standardized methods were indicated as superior to individually tailored approaches (Fals-Stewart et al., 1993; Lindsay et al., 1997). Confounding variables, such as the therapeutic relationship, were not controlled in all research studies. However, when this did occur, standardized methods still produced superior results (Chambless & Ollendick, 2001). Chambliss & Ollendick (2002) also investigated generalization of EBI laboratory research. Despite the small body of research reviewed, they found that EBIs produced similar positive results in both research and clinical settings. However, proper training for psychologists and the selection of an appropriate client population contributed to intervention effectiveness.

Identifying the most effective approaches was considered by the task force as effective in reducing the potential for practitioner bias. Kendall (1998) asserts that practitioners' selection of interventions is often more based on clinical judgment than it is the empirical evidence of the method. When causal relationships for improved outcomes are not examined, inaccurate attributions for success may occur. Implementation of EBIs increases the potential for outcomes that are strongly linked to treatment methods. Henry (1998) also supports the use of manuals for reducing the reliance on clinical judgment. He indicates that manuals, when used in the proper context, can guide clinician's application of therapeutic techniques and facilitate integrity of application.

In conclusion, Chambless and Ollendick (2001) predict that economic and societal pressures for accountability will sustain the momentum of interest in evaluating clinical psychology interventions. Criteria for establishing the empirical basis of studies by the Division 12 Task Force represent an initial step in ensuring support for the most efficacious treatments, both by practitioners and by managed care companies. However, quality mental health services require more than manualized procedures. For psychotherapy practice to be effective, practitioners must be competent in a variety of therapeutic models and skilled in fostering client relationships. Thus, EBIs are only one ingredient in developing an effective approach to clinical practice.

Evidence-Based Interventions in School Psychology

To meet the unique challenges of service delivery in schools, a separate but similar movement in the field of school psychology began in the late 1990s. The Evidence-based Intervention Task Force, founded in association with the Division 16 of the American Psychological Association and the Society for the Study of School Psychology (SSSP), was originally created for the purpose of evaluating the empirical basis of interventions for application to schools (Kratochwill & Stoiber, 2002). More recently, this coalition of practitioners, trainers and researchers has refocused its efforts promoting the integration of EBIs into practice settings. This objective comprises the ambitious goals of improving the quality of school psychology research and narrowing the long standing gap between those two domains.

The need to implement effective school-based interventions has been promoted by concerns about falling education statistics. More recently, the academic and behavioral problems of children have received national attention. Many children continue to perform poorly on standardized tests including the

ACT, SAT, and other state scholastic aptitude assessments. Reading and writing skills have fallen below achievement standards in many states with the lowest levels of comprehension in low income school districts. In addition, it has been estimated that approximately 20% percent of school children experience a variety of behavioral and social/emotional difficulties that render them at- risk of school failure. It has also been documented that most of these children do not receive appropriate intervention services (U. S. Department of Health and Human Services, 1999).

Scientifically-based methods are increasingly employed by practicing school psychologists to foster more positive outcomes in student performance. For example, the NASP publication, *Interventions for Academic & Behavioral problems I: Preventive and Remedial Approaches* (Shinn, Walker & Stoner, 1991), was revised and expanded in 2002 to enhance the efficacy of daily practice. Additionally, federal regulations, such as those included in NCLB and revisions of IDEA, specify the use of strategies with empirical support for producing increased student achievement. Other examples characterize the growing interest in a scientific orientation for educational and psychological practice: the National Center to Improve the Tools of Educators (Carnine, 2000) and the adoption of performance-based guidelines by the Office of Special Education (Christenson et al., 2002).

The research community has also contributed to this effort. A number of meta-analytic studies have been conducted to report on the overall effectiveness of school-based interventions for children with learning disabilities (e.g., Haney

& Durlak, 1998; Kroesbergen et al, 2003). Research in this area is directed at discerning the efficacy of interventions conducted in laboratory settings. However, these studies provide little guidance to practitioners in applied settings. Consequently, there was a definitive need for the identification of evidence-based interventions for school-age children and the efficient dissemination of this information to practitioners.

Currently, a number of organizations have begun efforts to identify effective interventions for application to school and other prevention settings. One such organization is the What Works Clearinghouse, which is supported by the U.S. Department of Education. The mission of the What Works Clearinghouse (WWC) is to identify effective intervention, prevention and educational programs that are of particular relevance to the education community (What Works Clearing House, 2004). Thus, the promulgated outcomes will be very relevant to the field of school psychology. The substantial financial and professional resources utilized by this organization have the potential for evaluation of intervention studies that is unmatched in the field. The goals of the WWC were closely aligned to those embraced by the Task Force. As such, these events had important implications for the direction of future work by the Task Force.

The Task Force and Its Mission

The Task Force on Evidence-Based Interventions reflects the expertise of leaders from school psychology. Dr. Thomas R. Kratochwill, a recognized leader in the field of school psychology, has consistently served chair of the group, with Dr. Karen Stoiber serving until 2004 and more recently, Dr. Kimberly Hoagwood appointed as co-chair. The Task Force was initially formed to evaluate psychological and educational interventions for school-age children. Since its inception, the group has developed guidelines for coding intervention studies into *The Procedural and Coding Manual for Review of Evidence-Based Interventions* (The Task Force on Evidence-Based Interventions in School Psychology, 2003) and conducted studies of Manual application. To date, the Task Force has published two version of the Procedural and Coding Manual. The original version of the Manual (Kratochwill & Stoiber, 2002) was revised by the Task Force in 2003.

Psychologists have also conducted in-depth reviews of the Manual. Many of these reviewers have noted challenges associated with the Manual criteria. Among their concerns, psychologists noted that study coding required extraordinary expertise in research methodology and a considerable investment in time and effort (Christenson et al, 2002; Nelson & Epstein, 2002; Prevatt & Kelly 2004).

The Task Force has responded to these challenges and the advent similar coding initiatives by reorienting the focus of their mission. Currently, the primary mission of the Task Force focuses on improving the quality of research training and providing professionals in schools with the knowledge and tools necessary to implement EBIs in practice. The organization is also considering theoretical models of change to promote broader use of EBIs by the field (The Evidence-

Based Intervention Work Group, 2005). Ultimately, the efforts of the Task Force are still directed at addressing the long standing research-practice gap. Advocating for the development of practice guidelines and identifying methods to facilitate adoption of EBIs into practice are currently being employed to meet this objective (Kratchowill & Shernoff, 2004; The Evidence-Based Intervention Work Group, 2005).

The task of integrating EBIs in practice raises new challenges for the Task Force. For example, the number of organizations involved in the coding process has resulted in the formulation of various coding standards. This diversity of efforts has led to uncertainty for practitioners responsible for interpreting coding results. In addition, few efforts have been made to tailor scientific research to meet the demands of practical settings and busy routines of school psychologists (Kratochwill & Shernoff, 2004).

Strategies to Guide the Implementation of EBIs

The volumes of intervention research available in the literature are rarely accessed by practitioners in school psychology. A number of reasons may account for the gap between research and practice. For example, the traditional methods by which research studies are drafted and disseminated have not been tailored to meet the unique needs of consumers. Practitioners have limited time in which to access and interpret research. Furthermore, they may lack the knowledge of research methodology necessary for this task (Kratchwill &

Shernoff, 2004). In addition, teachers who are often responsible for intervention implementation may require training in EBI implementation.

In response to these concerns, the Task Force has identified five strategies for the purpose of advancing evidence-based practice. These include: a) developing a practice-research network, b) promoting an expanded methodology for evidence-based practices; c) establishing guidelines for implementing EBIs into practice, d) creating professional development opportunities and e) forging a partnership with other professional groups (Kratchwill & Shernoff, 2004).

One of the main objectives of the Task Force centers on developing a dually beneficial relationship between research and practice. Recommendations for creating this synergy have come largely from the research community and particularly from members of the Task Force. Inroads toward achieving this goal are already reflected in the diverse composition of the Task Force. Comprised of both researchers, practitioners and trainers, it serves as a model of reciprocal practice.

In this model, practitioners participate as part of the research team and provide information pertaining to a variety of intervention components. Of particular interest to research teams, practitioners often possess a unique perspective on intervention contexts (Kratochwill, & Shernoff, 2004). Although often difficult to acquire, details pertaining to how the intervention actual "works" in practice are also obtained through these consultations (Kratochwill, 2002). As part of the Practice-Research Network, practitioners are able to educate researchers about a variety of variables related to the school culture that may impact intervention outcomes. Diversity, the school climate, and the quality of instruction are just a few of these variables. This alliance also offers reciprocal benefits. During collaborations, researchers can provide practitioners with detailed descriptions of intervention procedures and communicate details that are critical to applications in practical settings.

A second strategy focuses on promoting research on EBIs. As part of the research agenda, Kratochwill and Shernoff (2004) emphasize four types of research design that will advance evidence-based practice. These include: a) efficacy studies measuring the effectiveness of interventions in controlled laboratory settings, b) transportability research used to evaluate generalization of intervention effects, c) dissemination research examining intervention agents use of protocols and d) system evaluation studies which measure school-wide implementation of intervention strategies (Kratochwill & Shernoff, 2004). Future efforts of the Task Force will center on endorsing this new conceptual framework for the field of school psychology.

Developing guidelines to facilitate implementation of EBIs interventions into practice is another idea that has been embraced by the Task Force. As mentioned earlier in this report, many obstacles constrain the efficient integration of EBIs in practical settings (Kratochwill & Shernoff, 2004). Guidelines will be developed by the Task Force for the purpose of educating trainers, graduate students and practicing school psychologists in strategies that facilitate the everyday use of EBIs. At the same time, the guidelines are intended to address psychologists' concerns regarding the inflexibility of manuals and impracticality of practice guidelines (Kratochwill & Shernoff, 2004).

Kratochwill (2002) also sees the task of creating guidelines with utility and flexibility in applied settings as a salient issue. Christenson et al. (2002) also support the role of clinical judgment in practice. They acknowledge that practitioners, particularly those trained in scientist-practitioner model, have expertise in conducting interventions in a school context.

In their recommendation for future initiatives of the Task Force, Nelson and Epstein (2002) call for comprehensive training of the coding process. The Task Force has responded to this need by adding coding heuristics and formulas to the Manual (Stiober, 2002). In addition, *the Coding Workbook for Evidence-Based Interventions in School Psychology* providing instructions on reliably coding single- and group-based studies has been created. The Workbook supplements information in the Manual by providing consumers with illustrations of the coding criteria. As a self-instructional tool, the Workbook has outstanding potential for use in training and practice settings. If effective, the Workbook could be instrumental in the preparation of school psychologists with a scientist-practitioner orientation. However, prior to implementing the Workbook, the Task Force will need to obtain research documenting its usefulness.

In the future, the Task Force argued that training must be provided to a variety of constituencies throughout school psychology including practitioners, researchers, scholars, trainers and graduate students. Current plans for professional development comprise an emphasis on competent selection of strategies and supervision of EBI implementation (Kratochwill & Shernoff (2004). It is hoped that training opportunities for practitioners will be offered through state, regional and national conferences. Professional organizations may also disseminate information about EBIs to the field. Collaborative effects with educators will be directed at increasing coursework in EBIs offered through APA-approved graduate programs (Kratochwill & Shernoff (2004).

Finally, the development of collaborative relationships with organizations responsible for evaluating EBIs is essential to the success of the project. Currently, there are ten organizations involved in the process of coding psychological and educational interventions (Kratochwill & Shernoff, 2004*)*. Task Force chairs recommend that dialog between these professionals focus on understanding each group's vision for the EBI movement. Additionally, chairs have also called for an examination of the various coding systems implemented by groups. Finally, collaborative efforts must also promote clear and consistent communication of coding results.

Rationale for Establishing an Evidence-Based Practice

The movement to establish evidence-based practice in school psychology is important for a number of reasons. Knowledge of evidence-based interventions is needed to effectively address the academic and behavioral problems of school-age children. In addition, Task Force efforts will enable school psychologists to implement EBIs as methods to meet the current challenge of service delivery in schools. The many functions served by school psychologists, which include assessment, direct and indirect intervention and counseling, represent just one of these challenges (Stoiber & Kratochwill, 2000).

Skillful implementation of EBIs is particularly relevant to effective consultation. Many challenges are associated with indirect service delivery (Kazdin & Weitz, 1998). Some of these issues include establishing a collaborative relationship with teachers, intervention integrity and consistent data collection. Students exhibiting problems with attendance, academic performance and on-task behavior may be referred to consultation teams or the school psychologist who must identify presenting concerns and implement effective interventions in short order. However, time and resources constraints may serve as barriers to the reference of scientific research in the process of designing efficient intervention approaches. Thus, training and support in the implementation of EBIs is necessary to the school psychologists' ability to meet daily practice demands.

The mission of the Task Force is focused on increased competency for practicing school psychologists. Thus, efforts of the Task Force are directed at providing psychologists with the skills and understanding necessary for implementation of EBIs in everyday practice. Benefits resulting from this effort include enhanced ability to: a) properly match interventions to specific student concerns, b) skillfully supervise intervention implementation, and c) promote enhanced student outcomes (Kratochwill & Shernoff, 2004). With both access to a variety of intervention and prevention approaches and increased knowledge of research methodology, practitioners' will be able to weigh the relative merits of interventions and select the most appropriate treatments. In addition to the potential for improved rates of program success, this practice may lead to greater roles for school psychologists in promoting a student's ability to achieve and surpass federal standards.

Another advantage of the evidence-based movement relates to an increased integration of current intervention research into the practice of school psychology (Stoiber & Kratochwill, 2000; Stoiber & Waas, 2002). Because of difficulties encountered with identifying and interpreting scientific research, practitioners may too often rely on experience to guide practice decisions. This has resulted in infrequent reference to academic research and a hiatus between the scientific and practice communities (Stoiber & Waas, 2002). In addition, the busy schedules of school psychologists often limit their ability to reference intervention research. Strategies implemented by the Task Force are intended to assist practitioners with interpreting scientific research and to promote their confidence in applying EBIs to practice while still meeting the confines of a demanding school schedule.

Just as randomized designs have been established as the "gold standard" of empirical research in other fields, the thrust to establish EBIs as common practice in school psychology denotes the movement towards a similar milestone in our field. It represents an opportunity to improve mental health service delivery for school-age children. It also holds promise for narrowing the current research-practice gap. Furthermore, integration of the EBIs in schools and clinical setting may promote a higher standard for service delivery in these settings.

The Procedural and Coding Manual

The Task Force accomplished the first of its goals by formalizing guidelines for the procedure of evaluating intervention studies into *The Procedural and Coding Manual for Review of Evidence-based Interventions* (Kratochwill & Stiober, 2003). The Procedural and Coding Manual has been designed to evaluate interventions within the five domains representing the wide variety of contexts that exist in schools. It has been divided into four sections that reflect different design qualities: a) group-based design, b) single-participant design, c) qualitative research methodology, and d) confirmatory program evaluation (Kratochwill & Stiober, 2003).

Directions for reviewing and coding prevention and intervention programs are presented in a similar format within each section. Initially, intervention studies are evaluated according to general characteristics of the theoretical and empirical underpinning, general design quality and statistical application. The second set of criteria, key evidence components or Key Features, examines internal validity criteria and other features important to school- or field-based implementation. The third type of coding considerations expands on numerical ratings in previous sections by providing descriptive information. This section includes detailed demographic information about the participants as well as the criteria used for their inclusion in the study. It addresses other external validation and feasibility indicators such as descriptions of program implementers and the intervention environment. Readers are supplied with this information so they may make judgments about the appropriateness of the intervention in relation to the specific needs of their target population (Kratochwill & Stoiber, 2002). Thus, both quantitative and qualitative criteria will be used to determine the level of intervention effectiveness. Each area will be assessed in relation to a 4-point scale with a “3” indicating strong evidence, a “2” designating promising evidence, a “1” equal to marginal or weak support, and “0” specifying a lack of evidence. Criteria in all areas include consideration of internal and external variables and analysis of the intervention environment (Kratochwill & Stoiber, 2002).

Critiques of the Procedural and Coding Manual

The Task Force has petitioned for an examination of the identification process and welcomed commentary from both within and outside the field of school psychology. Psychologists have responded with both praise and criticism. Some have applauded the Task Force for taking an important step toward aligning the research standards in school psychology with that in other branches of the psychology (Christenson et al., 2002; Levin, 2002). These individuals view the endeavor as establishing a “gold standard” for school psychology while at the same time creating a method that is tailored to studies in school or clinical settings.

The movement has also been subject to criticism (Christenson, et al., 2002; Nelson & Epstein, 2002; Waas, 2002). Current concerns relate to both the coding process and dissemination of results. While recognizing these limitations, the Task Force responded by providing evidence to support their decisions and defining directions for the future. Despite any shortcomings of the process, school psychologists appear to agree the EBI movement represents an unparalleled opportunity to advance research and service delivery in school psychology.

Strengths of the Manual

Much of the controversy surrounding the Task Force has centered on its most notable publication, the *Procedural and Coding Manual for Review of Evidence-Based Interventions* (The Task Force on Evidence-Based Interventions in School Psychology, 2003). The Manual employs a dimensional approach to coding of a wide variety of study criteria. The intention of the Task Force was to develop a comprehensive coding scheme that was useful in clinical, research and university settings. By adopting the approach, the Task Force in School Psychology attempted to address criticisms that were encountered by Division 12 of the APA: a) it provides comprehensive and informative information; b) it rates the factors for scientific evidence of interventions on a number of dimensions; and c) it promotes practitioner's ability to use clinical judgment in applying interventions to specific clients (Waas, 2002).

Levin (2002) points out other assets associated with the Procedural and Coding Manual. First, the Manual categorizes studies according to type of research design. He believes that distinctions incorporated in the structure supply the Manual user with essential information about the intervention contexts. More importantly, the Task Force has accomplished the task of developing the most comprehensive resource for empirical research validation to date. Through the identification of general methodological qualities and statistical procedures required to evaluate study effectiveness, the Manual provides a standard for both researchers and practitioners. If used by graduate students or practitioners, it can function as a reference for essential components of scientific research.

Other strengths of the Manual comprise a focus on distinguishing intervention outcomes and contexts. Levin (2002) points out the inclusion of "key" and "ancillary" outcomes as an exceptional feature of the Procedural and Coding Manual. He credits this feature with promoting Manual user's ability to link specific outcomes with particular intervention components. Among other notable characteristics is the Manual's distinction of educational and clinical significance (Levin, 2002). While encouraging the consideration of factors "internal" to interventions, such as participant characteristics, it also advances the examination of "external" factors relating to educational consequences.

A focus on factors of participant selection and attrition is another strength of the publication. According to established standards of scientific research, this feature is critical to the establishing the validity of comparing randomly assigned study groups (Levin, 2002). Levin notes, however, that developing a clear understanding of attrition figures warrants an investigation of lost participant demographics. This approach would include expanding Manual criteria beyond mere numeric documentation to fully examine the scope and implication of this factor. Finally, Levin (2002) considers the criteria for study replication as one of the Manual's most prominent assets.

Shortcomings of the Manual

Despite its many advantages, numerous challenges have surfaced in reviews of the Procedural and Coding Manual. Prior to coding studies, a reviewer must conduct a comprehensive review of relevant intervention literature. Next, there is an extensive article coding process. At this stage, reviewers may dedicate a number of hours to the task of navigating the comprehensive and complex coding criteria. For example, the evaluation of a group-based design intervention would require the consideration of 3 major categories and 24 subcategories, which also include multiple rating criteria.

The comprehensive coding system has resulted in other problems for evaluators. An initial challenge arises when information required for coding is not included in research studies (Christenson et al. 2002). Reviewers often encountered cases where sufficient information for calculation of statistical measures is absent from study methodology. Data for calculating effect size is just one example of this concept. The task of contacting researchers to verify additional information has proven to be a time-consuming process that is sometimes unsuccessful (Levin 2002). In the absence of this relevant information, evaluators must base their decisions on judgment or inference. This occurrence presents challenges to the reliability of the coding process (Prevatt & Kelly, 2004).

Still other challenges existed with regard to coding outcomes in relation to reviewer's knowledge of and experience with scientific research. In their experience, Christenson, et al., (2002) found that accurate coding of studies requires considerable knowledge of the literature within a given research domain as well as competency in research methodology. They point out that reviewers who lack this background may fail to identify studies of primary importance or have insufficient knowledge of the theoretical/empirical basis of a particular research domain. The authors also note that advanced graduate students and professionals trained in a scientist-practitioner oriented program may be best suited for evaluating studies.

A review of dropout prevention and school completion by Prevatt & Kelly (2004) indicates that difficulties may be encountered by psychologists at all levels of training. As part of that study, two faculty members and two Master's level graduate students coded 18 studies. Difficulties experienced by the graduate students demonstrated that they lacked the methodological and statistical background to conduct a valid review of the studies. In addition, considerable discussion and interpretation of statistical procedures was necessary for faculty members to perform the evaluation. In some instances, issues of inter-rater reliability arose when reviewers made inferences about vague coding criteria. To better verify coding results, the faculty members conducted an independent review of the articles and compared the results. Multiple reviews and consultations were required for researchers to produce comparable ratings and develop a confidence in their results.

Christenson et al. (2002) state that problems with reliability issues may occur when individuals employ different coding approaches. For example, variability in outcomes often arise when one evaluator closely considers intervention context while another reviewer adheres more strictly to the Manual criteria and ignores contextual factors. In light of the contribution of inter-rater reliability to the overall validity of Manual, this remains an important issue with the coding criteria.

The time and effort involved in the study-coding process is a challenge faced by both experienced researchers and advanced graduate students. Prevatt and Kelly (2004) estimate the time invested in their review of dropout prevention and school completion research was approximately 80-90 hours. As a result, they concluded that this combined time and effort would prohibit broad use of the Procedural and Coding Manual by school psychologists. The state of Hawaii used the Manual to conduct a similar analysis of treatments for childhood disorders (Chorpita, 2002, as cited in Nelson & Epstein, 2002). In this study, it took 20 individuals 9 months to review 115 studies. Based on the time restraints experienced by the researchers, Nelson and Epstein (2002) argue that this limitation constrains the potential for similar research by individual psychologists or school districts.

Conducting a literature search that yields a representative study sample was another problem that arose in reviews conducted by both novice and experienced researchers. Nelson and Epstein (2002) report that despite advances in computer technology, identifying relevant literature is more difficult today than it was 15 years ago. Publication bias and the proliferation of databases with different search perimeters contribute to this problem. The ability to conduct effective searches of the literature requires that the researcher gain detailed knowledge of the search procedures, which may require training from a librarian or research specialist.

The task of identifying specific search terms has been a particular challenge. This is an important step that provides a framework for the search and a basis for defending the inclusion or exclusion of particular studies (Prevatt & Kelly, 2004). Guidelines for conducting searches of the literature and identifying appropriate search terms that have been added to the Procedural and Coding Manual address this issue.

Selective reporting by research journals also adds to the difficulty of locating an adequate and representative sample of studies (Nelson & Epstein, 2002). This problem also leads to less accurate evaluations of overall program effectiveness. Therefore, the criteria must also include methods to detect and eliminate publication bias.

Another disadvantage of the Manual stems from the absence of clear directions for the coding process. Specifically, the Manual fails to provide evaluators with a step-by-step process that promotes adherence to both objectivity and methodological rigor. Nelson and Epstein (2002) recommend a procedure that starts with the review of a selected research domain by a qualified panel. This method should be combined with a well-defined literature search process.

Next, a process for synthesizing individual research studies and generating a related report needs to be developed to as a final step to study coding.

A challenge once faced by the Task Force is that of translating the coding criteria into a practical instrument for practitioners. The format used by the Task Force to disseminate coding results also had implications for its utility in school, university and clinical settings. The coding protocol includes a chart for summarizing study evidence that indicates both numeric and descriptive ratings. While some have suggested that the structure lacks necessary depth (Waas, 2002) others have criticized the evaluative criteria in the Manual as too extensive (Durlak, 2002). If the coding system is deemed as too expansive and "overwhelming" for the practitioner, it may be subject to further consolidation. However, a condensed format of the Manual may exclude details about intervention conceptualization, implementation, and outcomes necessary to enable informed practice decisions (Waas, 2002). The degree to which the coding schema summarizes one study, or more importantly, a synthesis of studies, was considered a factor that would influence its use by busy school psychologists. However, this factor also influenced the amount of information that was included.

Problems with establishing uniform standards for presentation of results emerged with reviews that utilized descriptive reporting of coding outcomes. In their review of the Fast Track Program, Lewis-Snyder, Stoiber and Kratochwill (2002) supplemented the coding results with commentary about, and excerpts from, the actual study (Waas, 2002). Since the purpose of the article was to illustrate coding of a group-based study, it is likely that additional information was included by authors to aid readers in an introduction to the process. However, this raises the following questions: a) is this additional descriptive commentary necessary to convey information about criteria ratings to consumers of EBI research; b) if so, how can these descriptions be uniformly developed? Both of these issues have implications for the utility of ultimate product. Reviewers maintain that coding results must be easily interpreted by the consumer but also convey evidence of empirical basis. When information is presented by Likert scale-type ratings, descriptive information, even when succinct, may expand upon the clinician's knowledge of the methodological rigor of research (Waas, 2002).

An additional concern regarded the presentation of reviews. Given the complexity of the Procedural and Coding Manual, the task of generating a final report to synthesize coding results for multiple studies was a formidable challenge. The review of 18 dropout prevention studies by Prevatt and Kelly (2004) generated 65 pages of data, which to meet publishing requirements, had to be reduced into a 5-page table. The researchers noted a number of difficulties associated with condensing this vast amount of information into a journal manuscript. Waas (2002) suggested that rather than presenting comprehensive information, succinct coding schemes might function as a primary resource for the consumer. Through this system, readers could initially judge the applicability of program for their particular concern. Subsequent references to original studies would further insight about the theoretical basis and methodology of the intervention.

Criticism directed at the school psychology coding criteria also concerned its lack of emphasis on contextual intervention variables. Rather than conceptualizing a school as comprised of teachers, students and buildings, it is more accurate to view each school as having it own unique culture with norms, rules and values that reflect the particular population of that school (Sarason, 1996). The effectiveness and ongoing maintenance of an intervention depends on its "fit" to the ecology of the setting (Lentz, 1996). Thus, the intervention context is an essential factor to identify when evaluating the merit of interventions. The cultural diversity of the client population and cultural competency of the practitioner also play important roles in determining intervention outcomes. A number of researchers have found that consideration of cultural, ecological, sociolinguistic and phenomenological backgrounds of consultees and clients are essential to the implementation of strong and culturally sensitive interventions (Behring & Ingraham, 1998; Ramirez, Lepage, Kratochwill & Duffy, 1998). The identification of evidence for empirical basis must include consideration of research design in terms of its applicability to the school environment and cultural sensitivity to ethnic minorities (Sue, 1999). This point also argues for inclusion of evaluative criteria for the variety of research-based designs conducted in schools. Christenson et al. (2002) suggest that the field positively embrace quasi-experimental designs along with purely experimental research to facilitate the examination of these critical external variables.

Hughes (2000) also advances the argument that evaluating the effectiveness of interventions goes beyond the identification of its empirical basis.

Hughes asserts that efficacy, demonstrated in laboratory research, is not sufficient for assuring utility of methods in practical settings. Thus, constructing a knowledge base useful to clinical practice involves a full investigation of change mechanisms. This requires examination of three dynamic factors: the setting, the client and the therapist. Rather than simply a property of the intervention, effectiveness should be viewed as comprising the distributed effects of these dynamic factors. "The view that effectiveness is a property of the intervention ignores that fact that children are embedded in multiple interlocking systems and that change is one system that may play out in different ways depending on both characteristics of the child and characteristics of the child's social ecology (Hughes, 2000). Therefore, the current "gold standard" criteria are too narrow to accommodate the successful transportation of EBIs into practice. Instead, interventions must be evaluated in light of a cultural perspective with an emphasis on risk factors and developmental concerns.

Problems may arise when well-established theories are supported in the literature but fail to meet EBI coding criteria (Hughes, 2000). Attachment Theory approaches are one example of such methods. The extensive body of literature that supports this theoretical perspective shows that investigation in random clinical trials does not equate to evidence of intervention effectiveness. Therefore, the author advocates for realistic practice guidelines predicated on an understanding of all factors relevant to clinical utility (Hughes, 2000).

Finally, criticisms have emerged with regard to the current organization of the Procedural and Coding Manual. In reference to this issue, Kelly and Prevatt

(2004) noted that the Manual structure organized by class of research design may generate more confusion than clarity. This difficulty for researchers is created by the disparity between the focus of each domain. For example, the School-Wide and Classroom-Based domain refers to the environment in which the intervention is conducted; the Family Intervention domain refers to the target of the intervention. As a result of these differences, the Procedural and Coding Manual may lack the specificity to ensure that the evidence base for the intervention is evaluated in relation to the problem identified by the study (Prevatt & Kelly, 2004).

Statistical and Other Related Issues

A number of points have emerged as central to establishing a valid conceptual framework for the Procedural and Coding Manual. First, the issue of whether experimental design should be adopted as the hallmark for establishing the effectiveness of interventions in schools has been a pervasive theme in the commentary of reviewers (Christenson, et al., 2002; Nelson & Epstein, 2002; Waas, 2002). These criticisms center on the omission of a defined set of criteria that indicate what interventions should be included or excluded from the label of "evidence-based". Critiques of the Manual have made comparisons of the current standard identified by the Task Force to the "gold standard" in clinical psychology. True experimental design was the preferred as the criterion in clinical psychology because of the cause-and-effect relationship that could be established.

However, Christenson et al. (2002) note that randomization is difficult to accomplish in school settings and therefore many current school-based studies fail to meet these rigorous standards. Furthermore, more contemporary studies, as opposed to those conducted 20 to 30 years ago, are more likely to meet the new methodology criteria. Therefore, using this measurement standard is likely to exclude a number of studies from examination by the Task Force. The authors also note that these disqualified programs may otherwise have utility in practical settings. Thus, if the Procedural and Coding Manual incorporates a paradigm emphasizing randomized design criteria, it will conflict with the realities of the school environment and a scientist-practitioner orientation (Christenson et al. 2002). Waas (2002) posits that "It will be important for EBI efforts in school psychology to accommodate the diversity of intervention objectives, participants, and methodologies in the evaluation criteria if these efforts are to narrow the continuing hiatus between research and practice". A comprehensive approach reflecting a variety of experimental research designs has in fact been developed by the Task Force. The current Manual includes acceptable criteria for studies that utilize randomized and nonrandomized (hierarchical and block) designs. Specific criteria for Quasi-experimental designs have also been addressed.

A second point refers to thresholds for key outcome criteria. Durlak (2002) argues that the Manual fails to distinguish between acceptable levels for evidence of intervention outcomes. He notes that the Procedural and Coding Manual, in its current form, indicates that outcomes merely "represent a valid and appropriate indicator" for the program under consideration.

While some critics of the Procedural and Coding Manual propose that too little emphasis has been placed on effect size, others applaud the Task Force for the importance afforded to statistical considerations. Levin (2002) has found that the Manual reflects an appropriate balance between statistical significance and the magnitude of an intervention. He points to the prevalence of single participant studies in field-based research as a reason to place equal emphasis on these factors. In contrast, Durlak (2002) argues that more weight should be placed on effect size because it provides information that is critical to understanding the impact of the intervention. He argues that because statistical significance figures fail to indicate the change realized from implementation of a particular method, these indicators bear less relevance on intervention evaluation.

Statistically insignificant outcomes may also coexist in studies with powerful effective sizes. Durlak (2002) illustrates this point with a study that produced increased graduation rates of 15%. These outcomes failed to be statistically significant at the 0.05 level. At the same time, this study also yielded an effect size of up to 0.40. School-based studies that often comprise small participant samples are also likely to produce less significant results. Therefore, improvements to the criterion levels for effect size were recommended. Currently, effect size magnitudes are measured in terms of outcome levels (i.e. 0.20, 0.50 or 0.80). In contrast, Durlack (2002) proposes that study effect sizes should be judged in relation to effect sizes produced by similar studies.

Examination of effect size in reference to contextual factors will also provide information about the clinical significance of research outcomes (Durlak, 2002). A study that generated a 0.20 effective size may seem insignificant until compared to similar studies or viewed from the perspective of change to the school environment. However, despite the potential merits of this process, it may be impractical when embedded into a complex coding procedure. Levin (2002) cited these difficulties in association with using the Manual to calculate effect sizes of multilevel studies.

Third, the inclusion of qualitative criteria in the Manual by the Task Force has generated vigorous censure from some experts in the field. A coding category for studies using qualitative research methods can be found under General Characteristics in the Procedural and Coding Manual. This section requires coders to identify adherence to the theoretical-empirical basis of studies, procedures for ensuring coding consistency, and evidence of a progression from abstract concepts to empirical study exemplars (The Task Force on Evidence-Based Interventions in School Psychology, 2003). However, an evaluation of statistical data is not included as part of the qualitative considerations. Waas (2002) points out that "inclusion of qualitative reports as forms of "evidence" codified by the association risks blurring the distinction between decision making based on sound evidence and decisions based on anecdotal reports". The adoption of qualitative data as part of the evaluation process was characterized by Nelson and Epstein (2002) as a "liberal" approach to demonstrating the effectiveness of program interventions.

Waas (2002) raised this issue as a possible obstacle to the generalization of the EBI coding results to a variety of practical settings. Although the inclusion of qualitative criteria in the Manual expands its focus to include educational studies utilizing a quasi-experimental design, it also results in a less stringent assessment of studies. The decision to include this category suggests that the Manual lacks a defined standard for empirical validation. It also raises the question of what criteria should be included to accommodate the research conducted in school-based settings.

Other problems with the Coding and Procedural Manual have emerged in terms of categorizing primary study outcomes. To evaluate primary outcomes, coders must be able to identify the “ultimate goal(s)” of interventions (Manual, 2003). However, definitions for behavior and academic outcomes vary depending upon how the target problem is conceptualized by the researcher. A variety of methods might also be employed to measure outcomes. For example, the definition of bullying may differ across studies of the same intervention. Furthermore, each study may utilize a different outcome measure (Prevatt & Kelly, 2004). Prevatt and Kelly also point out that it may not be possible to identify the best practice approach given the multiple operational definitions and interventions methods that exist in the literature.

Reactions by the Task Force

The Task Force chairs have responded to commentary from the field by citing their rationale behind the purposes and structure of the Procedural and Coding Manual. However, the group also acknowledges the critical issues raised may shape the ongoing efforts of its endeavors. Concerns regarding the criteria used to establish empirical support for interventions in school psychology were addressed directly by the Task Force. Kratochwill (2002) agreed with the assertion by Christenson et al. (2002) related to the narrow scope of randomized design criteria. Thus, the Task force has promoted randomized experiments as the "gold standard" to which school psychology researchers should aspire to achieve. However, Kratochwill also pointed out that developing an understanding of "what works" for a particular concern necessitates consideration of research methods traditionally used in field-based studies. Stoiber (2002) describes uses a broad-based approach that includes: a) addressing contextual variables, b) incorporating a focus on scientific principals and c) promoting the sharing of evidence-based data in research and practice.

Kratochwill (2002) also acknowledged the concerns raised by Nelson and Epstein (2002). This issue relates to potential for unfavorable ratings that may arise from comparisons of school-based research to conventional standards for experimental design. Kratochwill (2002) asserts that inclusion of comprehensive coding criteria allows for a broader investigation of intervention research. This factor is critical when studies vary in terms of validity features.

The related decision made by the Task Force to incorporate qualitative methodologies into the Manual as a basis for evidence was also discussed. This issue was addressed by demonstrating that the rich and descriptive information found in qualitative studies can enhance external validity and aid practitioners in their efforts to apply scientific research to practice (Kratochwill & Shernoff, 2004). Stoiber (2002) asserted that "Our EBI coding scheme was intended to embrace different modes of inquiry so as to strike a balance between encouraging research rigor and building capacity to do field-based research and high-quality intervention practices".

Criticisms regarding an absence for consideration of contextual factors in the Manual are also addressed by the Task Force. Stoiber (2002) believes that Task Force efforts have gone beyond that of clinical psychology and other fields by extending the definition of effective methods to include the conditions and context of the intervention. Their knowledge of the difficulties that surround conventional experimental design studies conducted in schools has led the group to adopt a broad, dimensional approach for evaluating school-based research. This approach has also received the support of the National Research Council (NRC). The NRC defends research methodologies that reflect the unique conditions in schools and are best suited to inform the practitioners working in this environment (Stoiber, 2002).

Lewis-Synder at el. (2002) agree with the argument advanced by Hughes (2000). Both authors hold the view that knowledge of the theoretical and empirical basis of interventions is necessary for determining "what works" for a particular setting and condition. Furthermore, Lewis et al. concur that a strong conceptual understanding enables practitioners to appropriately modify intervention principals to accommodate the populations they serve. This ideal has been incorporated into the Procedural and Coding Manual criteria. When coding a research study, evaluators are required to consider both research methodology and theory as basis for empirical support. Thus, the coding scheme provides support for interventions based on established theoretical perspectives of a particular academic or behavior concern.

The Task Force has addressed the issue of cultural diversity by integrating principles from the APA *Guidelines for Providers for Psychological Services to Ethnic, Linguistic and Culturally Diverse Populations* into the coding criteria of each research domain (Kratochwill & Stoiber, 2002). However, evaluation of cultural factors may be more complex than originally anticipated by the Task Force. Wampold (2002) has found that the examination of culturally effective interventions require the identification of culture as a specific contextual variable. This factor, when combined with particular remediation methods, will produce positive outcomes for the specified client population. Adopting this point of view assumes a number of assumptions: a) specific intervention components correlated to specific outcomes, b) specific intervention ingredients may affect populations differently due to cultural or socioeconomic factors, c) the skill, sensitivity, knowledge and culture of program providers impact intervention results (Wampold, 2002).

Validity concerns have emerged with regard to the identification and review process. Problems with identifying adequate study samples are related to publication bias and difficulty with obtaining missing information from studies necessary for coding as noted by Levin, (2002). To address these study identification issues, the Task Force has added “confidence ratings” to the Manual. These ratings enable coders to express their degree of confidence in coding outcomes (Kratochwill, 2002). Missing or unavailable information can also be designated as “Not Reported” (Stoiber, 2002). Stoiber points out that detailed analyses are required for judging intervention effectiveness. This option permits reviewers to evaluate studies with methodologies that deviate slightly from Manual criteria.

A number of reviewers for the Manual have noted factors such as competence in search procedures, knowledge and experience of coding criteria, and representativeness of the research as potential threats to validity of the EBI process. In his address of these concerns, Kratochwill (2002) details a number of measures initiated by Task Force to deal effectively with these concerns. First, review of each research domain is conducted by a diverse panel of professionals. Second, the initial training for the coding process has been expanded to include a wider focus on the literature review process. This initiative has also led to the inclusion of more explicit directions to the Procedural and Coding Manual. In addition, a Coding Workbook containing illustrations of coding criteria have been developed to prepare school psychologists for using the Manual.

Revisions to the Manual have led to increased measures of its validity. As a result, high indicators of inter-rater agreement were reported by the Task Force. Stoiber (2002) argues that concordance rates averaging .85 were recorded in studies by field-based Task Force members and graduate students. Too ensure that reliability for subsequent literature reviews matches levels of agreement obtained during trials, procedures for inter-rater reliability have been added to the appendix of the Procedural and Coding Manual.

Challenges associated with the summary and communication of coding results has been noted by a variety of reviewers (Durlak, 2002, Nelson & Epstein, 2002, Wampold, 2002). Traditionally, narrative reviews are a used to summarize results. The Task Force has expanded this method to also comprise meta-analysis. The combination of procedures is used to arrive at casual inferences (Kratochwill, 2002). Hence, the Task Force has incorporated both approaches to the Manual while maintaining a preference for quantitative reviews that include effect size data (Kratochwill, 2002).

In response to concerns about the complexity of the coding criteria, Stoiber (2002) acknowledges that reviewers must consider extensive information. However, she disagrees with Nelson and Epstein's (2002) claim that time constraints prohibit the viability of the Procedural and Coding Manual in clinical settings. Although initial trials may take up to 5 hours, subsequent coding has resulted in significant practice effects. Stoiber (2002) has found that coding practice enables reviewers to conduct this process in 2 hours or less for a study.

Potential Outcomes of the EBI Movement

The work of the Task Force has the potential of advancing the field with integration of scientifically-based principles into every day practice. The outcome of the ambitious effort is highly anticipated by both school psychologists and other education professionals. However, when considering the history of the evidence-based movement, one recognizes that while the mission of the Task

Force is attainable, the group must overcome a number of obstacles in this journey.

A number of potential outcomes for the EBI movement are suggested by Waas (2002). In the first and least desirable of these outcomes, the EBI findings (which now will be furnished by groups outside of the Task Force) are largely ignored by school psychologists who view the recommendations as simply a list of exclusive interventions that are irrelevant to practice. Precautions taken to identify interventions based on a range of effectiveness may prevent the likelihood of this result. The second scenario has direct implications for the development of practice guidelines. Guidelines that require practitioners to apply EBI in a highly structured fashion will limit rather than enhance their strategic use of intervention programs. Thus, contrary to school psychologist's interest in occupying a greater role in schools, they are likely to be marginalized as deliverers of specified treatments.

In the third scenario, school psychologists, through the application of effective training and resources, assume a leadership role in the implementation of EBIs in schools. If provided with tools that inform their understanding of Manual, school psychologists will able to select interventions best suited to their objectives, student populations and setting requirements. As part of this scenario, the strategies implemented by the Task Force will need promote school psychologists' adoption of new approaches to practice (The Evidence-Based Intervention Work Group, 2005). Consultation provided to practitioner must also provide guidelines for effective application of EBI methods. Ultimately, school psychologists who practice with greater competency will be met with increased in client satisfaction. Confidence stemming from expertise in evidence-based intervention approaches will likely result in increased implementation of these methods. As such, the potential for bridging the current gap between research and practice may be realized.

Summary

In attempting to advance the field of school psychology, the Task Force has adopted ambitious goals: to establish criteria for the empirical basis of intervention programs and to strengthen the connection between research and practice in the field. This task has proved to be more time consuming and difficult than was initially anticipated. The number of obstacles in application of the Manual has prompted substantial criticism. Many of these issues remain unresolved with the finest scholars in our field still in the process of refining a "roadmap" for evidence-based practice (Gutkin, 2002).

Current challenges and other developments in the field caused members of the Task Force to re-evaluate their efforts and set new priorities for future goals. Consequently, the group has redefined its mission to include a focus on promoting broad use of EBIs in school psychology. As part of this strategy, the Task Force has proposed expanded training efforts and the application of models for adoption of innovations be integrated as part of the EBI movement.

The Task Force continues to face a number of challenges in its effort to expand the use of EBIs in practice environments. Most importantly, the Task

Force will need to implement a strategy for educating school psychologists about criteria in the Procedural and Coding Manual. The development of the Coding Workbook is essential to the execution of this initiative.

Despite these problems, the EBI movement is not without substantial merit. The accomplishments of the Task Force have been widely recognized by school psychologists and by those in related disciplines. Even those individuals who have raised concerns about the Manual concur that the movement holds outstanding potential for change in school psychology. Although development of the Manual has generated complex challenges, it is only by adherence to such rigorous standards that school psychological research will rise to levels attained by other disciplines. Thus, the work of the Task Force has the potential to improve the day-to-day practice of school psychology in meaningful ways.

CHAPTER 3

METHOD

Purpose of the Study

Research for the present project will focus on a case study of *the Coding Workbook for Evidence-Based Interventions*. The purpose of the study is to evaluate the effectiveness of the Workbook as a training instrument for the EBI coding process. In addition to coder's observations of the process, measures of inter-rater correspondence and consensus code agreement collected prior to and following implementation of the *Workbook* will be used to evaluate its utility. Study coding times will also be documented to examine possible variations in time requirements. The descriptive analysis generated by the study will potentially include recommendations for future development of the Coding Workbook.

Instruments

A number of instruments were used to collect data for this project. Instruments for coding responses include protocols contained in the Coding Workbook and the Manual. In addition, an observation log form was developed for reviewer commentary about the training and coding process.

The Procedural and Coding Manual and Protocol

The Procedural and Coding Manual for the Review of Evidence-Based Interventions (The Task Force on Evidence-Based Interventions in School Psychology, 2003) was used in this study. The Manual provides coding criteria and a protocol for coding responses. The protocol was used to record codes selected by reviewers for each category of Manual criteria. Inter-rater agreement was calculated based on these responses.

The conceptual framework for Manual has been presented by Kratochwill and Stoiber (2002). Studies are evaluated based on three factors: a) General Characteristics, b) Key Features and c) Descriptive or Supplemental Criteria. The first coding category identifies the type of research design and examines strength of the statistical or theoretical foundation of the study. In the second set of criteria, internal validity factors were assessed. These factors included a focus on outcome measurement procedures, group equivalency methods and the statistical significance of intervention outcomes. Implementation fidelity and replication were also evaluated. The third category addresses external validity indicators such as participant demographics and specific methods of implementation. For these criteria, reviewers used descriptive reporting to evaluate external and internal validity factors.

Reviewer decisions regarding Manual criteria were entered in the coding protocol for Group-Based Design studies. In addition to ratings for each coding category, tables were completed to analyze various components of study methodology. Descriptive information about study methodology was also documented. Reviewers' data entry also included the Summary of Evidence form. A copy of the Coding Protocol for Group-Based Designs (The Task Force for Evidence-Based Interventions in School Psychology, 2003) is available on request.

Coding Single-Participant and Group Design Studies: Coding Workbook for Evidence-Based Interventions

The Coding Workbook (Shernoff & Kratochwill, 2004) comprises a self-instructional training for coding research studies using criteria developed by the Task Force. Examples are provided for both single-participant and group-based participant research designs. Workbook instructions inform users to enter codes in the Manual protocols for Group-Based and Single-Participant Design studies. A copy of *the Coding Workbook for Evidence-Based Interventions* (The Task Force for Evidence-Based Interventions in School Psychology, 2003) is available on request.

The Observation Log Form

The observation log form was developed to record reviewer commentary about use of the Workbook and Manual. It was implemented during the case study to document reviewers' feedback about interpreting coding criteria and the issues that arose during this process. Observation logs were also used to record suggestions for improving criteria in the Workbook. Refer to Appendix F for a copy of the observation log form.

Materials

Results of the literature search produced a small sample of intervention research and four studies were randomly selected from this sample for Manual coding. Studies correspond to the Task Force domain of Task Force Domain of

School-Wide and Classroom-Based programs. Citations and abstracts for each study are presented below.

Larson, K. A. (1989). Task-related and interpersonal problem-solving training for increasing school success in high-risk young adolescents. *Rase, 10(5),* 32-41.

The study reviewed as article 1 investigated the efficacy of a task-related and interpersonal problem-solving intervention for "difficult-to-teach", low SES minority students. Through methods of random assignment to experimental and control groups, the study demonstrated that the intervention significantly improved report card grades and reduced misbehavior.

Seifer, R., Gouley, K., Miller, A. L., & Zakriski, A. (2004). Implementation of the PATHS curriculum in an urban elementary school. *Early Education and Development, 15(4),* 485.

A study of the PATHS curriculum implemented in an elementary school serving low income minority students was coded as article 2. The PATHS curriculum was shown to result in higher social-emotional competence for the intervention than the control group.

Kam, C., Greenberg, M. T., & Walls, C.T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4(1),* 55-63.

The study selected as article 3 examined implementation quality of the PATHS curriculum. The study sample comprised 350 first grade students in six urban public schools. The intervention was found to be effective in select schools for improving students' social-emotional competence and reducing aggressive behavior.

Greenberg, M.T., Kusche, C.A,, Cook, E. T., & Quamma, J. P. (1995). Promoting emotional competence in school-aged children: The effects of the PATHS curriculum. *Development and Psychopathology, 7,* 117-136.

The effectiveness of PATHS program for regular education and special needs students was investigated in article 4. The intervention field trial involved 286 students in grades 2 and 3. Results demonstrated improvements in emotional recognition and expression by both low- and high-risk student populations.

Case Study Reviewers

Research for this case study was conducted by three reviewers. The faculty reviewer was a professor of school psychology at the University of Maryland. Dr. Strein's distinguished career includes experience as a school psychologist, professor and researcher. His research interests include children's social-emotional learning and professional issues in school psychology.

The other two reviewers were graduate students in the University of Maryland School Psychology Program. The graduate students have obtained a similar level of training and research experience. The researcher was a Specialist level graduate student with two years of coursework and one year of internship in school psychology. In terms of research experience, the researcher has participated in Level of Implementation data analysis for Instruction Consultation Teams in Baltimore City Schools. The secondary reviewer was a pre-doctoral level graduate student with three years of coursework in school psychology. The secondary reviewer has two years of research experience that include one year as a graduate assistant in the University of Maryland Lab for Consultation Teams. Coursework completed by the graduate students that are specifically relevant to using the Coding Workbook include two courses in Quantitative Methods.

There a number of reasons for the inclusion of the secondary and faculty reviewer in this study. First, the secondary reviewer was asked to participate in the project based her current level of school psychology training. The participation of a graduate student with training similar to the researcher was needed to accomplish equivalent pre-and post-training comparisons of coding performance. In addition, the inclusion the graduate students and faculty as participants has implications for the generalizability of study results. Coding performance of the graduate students was assumed as comparable to other Specialist level graduate students in scientist-practitioner oriented programs. In addition, outcomes generated by the faculty reviewer were assumed as generalizable to other university faculty members with considerable research experience.

Pilot Study

The faculty reviewer and researcher conducted a trial study of Procedural and Coding Manual in the fall of 2003. Article 1 entitled *Task-Related and Interpersonal Problem-Solving Training for Increasing School Success in High-Risk Young Adolescents* (Larson, 1989) was coded during this initial study phase. Due to its examination of a social skills intervention, the study differed from PATHS research reviewed in later phases of this case study. Despite this difference, it shared a number of characteristics with other programs selected for this study. For example, it utilized a completely randomized design and well-defined research methodology. In addition, minority students of low income backgrounds served as study participants.

The faculty reviewer and researcher used the Procedural and Coding Manual to conduct an independent review of article 1. The Coding Workbook was not used for the pilot study. Results were compared following the review process. Overall, both reviewers found the initial experience of coding a study to be a cumbersome and time-consuming process. The graduate student noted that it required 6 hours and 45 minutes to complete one study. Although coding required a shorter amount of time for the research professor, this task still required 4 hours and 30 minutes for completion.

The comparison of coding outcomes prompted considerable discussion between the two evaluators. As a result of the analysis, an inter-rater agreement of 59% was established. Results of the pilot study also prompted dialogue about potential preparation methods to increase coding reliability.

The results of the pilot study demonstrated a need for evaluators to expand their current knowledge of the Procedural and Coding Manual. The Task Force designed the Coding Workbook specifically for this purpose. Thus, the present study will utilize the Workbook as preparation for the second coding trial.

The implementation of the Workbook as a training tool is based on a number of assumptions. One assumption is that Workbook explanations will enhance coders' understanding of the statistical properties and technical criteria in the Procedural and Coding Manual. A second assumption is that practice with Workbook examples will clarify technicalities of the coding process and its use will result in increased ability to apply Manual criteria. The Workbook should also serve as a model for coding research during the case study.

Overview of Procedures

Descriptions of procedures used in this study will begin with an overview of the case study design. It will include a brief description of methods for data collection and analysis. A detailed discussion of case study methods follows this introduction.

The present study consists of a pre-training coding pilot, a training module and a post-training measure of study coding. The pilot study coding of article 1 by the faculty reviewer and researcher served as the pre-training measure. As preparation for subsequent coding, the reviewers completed self-instructional training using exercises in the Coding Workbook. The faculty member, researcher and the secondary reviewer coded workbook exercises. The self-instructional training

During the third phase of the study, the Procedural and Coding Manual was used to code intervention studies. Studies were reviewed according to guidelines for the EBI coding process: a) use best practice in conducting a literature review; b) select appropriate coding criteria; c) complete coding sheets for each study; d) complete the summary coding form; e) summarize coding results; f) write the research report (Kratochwill & Stoiber, 2002). The faculty reviewer and researcher, who participated in the pilot study, coded article 2. The secondary reviewer, who did not participate in the pilot study, coded article 1. Phases of this research project are summarized in table 1.

Table 1

Phases of the Case Study

Study Phase	Faculty Reviewer	Researcher	Secondary Reviewer
(1) Study Coding	Article 1	Article 1	-
(2) Self-Training	Workbook	Workbook	Workbook
(3) Study Coding	Article 2	Article 2	Article 1
(4) Study Coding	-	Article 3 Article 4	-

Research studies include article 1(Larson, 1989), article 2(Seifer et al., 2004), article 3(Kam et al., 2003) and article 4 (Greenberg et al., 1995).

A variety of methods were used to measure effectiveness of the Coding Workbook. Pre-and post-training coding responses were compared to calculate inter-rater correspondence. Reviewer coding proficiency was measured through comparisons to codes determined by a consensus of researchers. Coding time for the pilot study and post-training review was documented. Research data also included extensive observation logs regarding the effectiveness of the training module and results of the study codings. Information obtained during the pilot study, training and coding trial was analyzed to determine study results. Data analysis included comparisons of inter-rater correspondence, consensus code agreement and coding time. An interpretation of observation logs about the

Manual and Workbook applications was also conducted.   Study results were reporting using a descriptive summary format. While addressing specific research questions, it also examines the degree to which the Workbook was found to increase reviewers' understanding of Manual criteria.

Procedures

Literature Search Process

A literature search focused on interventions in the Task Force Domain of School-Wide and Classroom-Based programs with disruptive classroom behavior as an area of specialization. The Promoting Alternative Thinking Skills (PATHS) Curriculum was selected as the program for post-training reviews with the Manual. The search utilized the following resources: (a) computerized educational and psychological databases, (b) the Dissertation Abstracts database, (c) published reviews and meta-analyses of interventions in the target domain, (d) the Science Citation Index, and (e) additional studies identified as relevant to the research domain (Kratochwill & Stoiber, 2002). Key words such as *Promoting Alternative Thinking Skills and PATHS* were used to conduct searches of computer databases such as *PsychInfo and ERIC*. If available, review articles were used to identify original studies for analyses.

Literature searches were conducted to identify a social skills intervention (Larson, 1989) for the pilot study (article 1). This procedure also identified a limited sample of the Promoting Alternative Thinking Skills (PATHS) program studies for the post-training measure. A small sample of three PATHS studies was randomly selected from results of the literature search. The studies included: a) article 2 (Seifer et al., 2004), b) article 3 (Kam et al., 2003), and c) article 4 (Greenberg et al., 1995). To be included in the sample, studies met the following inclusion criteria: a) the researcher(s) must present data on the use of the Promoting Alternative Thinking Strategies (Kusche & Greenburg, 1994), b) the study must be school-based, c) the research should be conducted with diverse student populations, d) the studies must by conducted in the past 20 years. Reviewed research included selected studies regardless of the significance of the findings.

Preparation of the Training Materials

A number of procedures were performed to prepare training materials. The Workbook developed by the Task Force includes a response protocol with pre-determined “correct” answers. For this study, responses on the Workbook protocol were eliminated and a separate answer key was created. As such, reviewers could then record their responses directly in the Workbook. Detailed directions reflecting this modified organization were developed. Instructions that differed from those originally presented in the Workbook advised reviewers to compare their responses to those indicated as correct by the Task Force and to make notations of coding accuracy. Reviewers were also specifically required to record feedback after coding each exercise. Other innovations included observation logs and detailed instructions for each phase of the case study. Instructions included also guidelines for Manual coding.

Self-Instructional Training using the Coding Workbook

The faculty reviewer, researcher and a secondary reviewer completed self-training exercises in the Coding Workbook. The purpose of this task was to measure the degree to which the Workbook exercises increase the coder's understanding of Manual criteria. Observation logs were used to document reviewers' feedback about the training process. Measured changes in inter-rater correspondence and consensus code agreement from the pilot study to post-training review were recorded to gauge effects of Workbook training. Methods for data collection and analysis are addressed below.

The self-instructional training was conducted by the three reviewers using materials and exercises in the Coding Workbook. The training module is organized according to the three strands of coding criteria in the Manual: Part I: General Characteristics; Part II: Key Features and Part III: Other Descriptive Criteria. Each part includes a study reference, identification of the Task Force Domain, a research study excerpt and applicable coding criteria. Each exercise has a coding protocol that corresponds to the relevant coding criteria.

To conduct Workbook training, reviewers were provided with instructions developed by the researcher. They were also advised to review the *Organization of the Coding Module* found on the first page of the Workbook. Instructions to reviewers indicated the materials and specific procedures for training. However, the Workbook coding was conducted independently by reviewers who could determine specific times and locations for training.

Materials required for training included the Coding Workbook, the Coding Workbook “response key”, the Manual and the observation log form. The Manual was used as a reference for information, such as formulas, that were not provided in the Workbook. In addition to dates and times for coding, the observation log form was used to record reviewer commentary. Two different protocols were used to record codes for Workbook exercises. Initially, it was planned that the reviewers would follow Workbook instructions and record their responses for exercises on a separate Manual protocol. However, since responses were removed from the Workbook, the faculty and secondary reviewer entered their responses on this protocol. Conversely, the researcher followed the Workbook instructions and recorded codes on the Manual protocol.

The training module was conducted using a 5-step process. Beginning with the General Characteristics exercise, reviewers read study excerpts and reviewed applicable coding criteria. Coding criteria for the Workbook generally corresponds to criteria in the Manual with some exclusions. Second, responses were entered in the Workbook protocol (or the Manual protocol by the researcher). Third, reviewer responses were compared with the “correct” answers in the “response key”. Fourth, selected responses were marked as “correct” or “incorrect” on the Workbook protocol. While correct responses signified a general understanding of the coding criteria, observation logs were used identify unclear aspects of the criteria. While checking their responses, reviewers also noted whether information on the answer key provided them with clarifications for miscodes. Finally, feedback about the process was recorded in observation logs.

Data Collection for Workbook Training

Two types of data were collected from the Workbook module. First, reviewers recorded responses to training exercises in the Workbook or Manual protocol. Second, observation logs recorded during training focused on observations and suggestions for improving the Workbook. The documentation form provided space for general comments about the Workbook criteria, assessment of information in the answer key, and suggestions for improvements to the exercise. Criteria easily understood and applied with confidence were identified. Specific issues and challenges of interpreting criteria in light of information provided in the Workbook were also noted. Workbook exercises without sufficient information to promote coder's understanding of the coding criteria or responses provided were emphasized. Reviewers noted clarifying information that was found in the Manual, but not the Workbook. Finally, reviewers wrote a summative evaluation of the Workbook. Opinions were documented regarding the effectiveness of the answer key for clarifying inaccurate responses, strengths and weakness of the Workbook and suggestions for additional exercises.

Intervention Coding with the Manual

Following completion of the Workbook, each reviewer conducted an independent study coding with the Manual. The faculty reviewer and researcher coded article 2 (Seifer, et al., 2004). The secondary reviewer, who did not participate in the pilot study, coded article 1. After the completing the peer-review process, the researcher coded articles 3 and 4.

The instruments for Manual coding varied somewhat from the training materials. The materials for this phase of the case study included: the Procedural and Coding Manual, the intervention study, the Coding Protocol for Group-Based Designs and the observation log form. The coding protocol was used to record coding selections and these results were added to the Summary of Evidence. Dates and times of coding sessions and feedback about the process were documented in observation logs.

All studies were coded according to instructions in the Coding and Procedural Manual. The Manual suggests review of studies within three general categories: (a) General Characteristics, (b) Key Features, and (c) Supplemental/Descriptive Information. Criteria in the first two categories were assessed using a 4-point scale with “3” indicating strong evidence, “2” designating promising evidence, “1” equal to marginal or weak support, and “0” specifying a lack of evidence. Decisions in the final area required identification of descriptive ratings.

Methodological qualities and statistical procedures of intervention studies were evaluated as part of coding General Characteristics. Evaluation in this area focuses on the appropriateness of the intervention to the school environment and the quality of behavioral assessment. The intervention was first be classified in terms of the type of research design (random or nonrandomized assignment).

Criteria for Statistical Treatment include examination of the Unit of Analysis, controls for familywise error and appropriateness of study samples. Reviewers also classified the type and stage of the intervention.

The Key Features category involves consideration of internal and external variables and analysis of the intervention environment (Kratochwill & Stoiber, 2002). Interventions were coded on the basis of nine key features. The Measurement criteria entail a determination of Reliability and Validity for primary outcomes measures. Next, the type of Comparison Group utilized in the study was identified. This category also includes criteria for counterbalancing of change agents, group equivalence and group mortality. As part of establishing Significance for Primary and Secondary Outcomes Significance, judgments about of appropriate analysis measures and calculation of study effect size were required. Evaluators then rated evidence for Educational Significance and Identifiable Intervention Components. These coding criteria also included evaluations of Implementation Fidelity, Replication and the Site of Implementation.

The final section documented descriptive information about intervention implementation. For Other Descriptive or Supplemental Criteria, reviewers indicated detailed demographic information about the participants as well as the criteria used for their inclusion in the study. Other feasibility indicators identified in the section included length or intensity of the intervention, characteristics of the implementer and intervention orientation. Information provided in this section is used to judge the appropriateness of the intervention for specific student population (Kratochwill & Stoiber, 2002).

Summary of Evidence

A Summary of Evidence form was completed for each research study. The summary includes overall ratings and a description of evidence for each indicator in the Procedural and Coding Manual.

Data Collection for Intervention Coding

Methods of data collection for Manual coding were similar to those used for the review of Workbook exercises. Responses were recorded on the Manual protocol. Observation logs were used to document detailed examinations of Manual coding. Similar to the process used for coding Workbook exercises, reviewers assessed the comprehensibility of the coding criteria and provided constructive feedback. Reviewers' commentary served as indication of their ability to apply Manual criteria following Workbook training. Observation logs also documented the time required to code each study. Finally, evaluators used descriptive reporting to appraise their coding performance and knowledge of criteria after using the Coding Workbook.

Inter-rater Correspondence

Coding results for the pre- and post-training study coding generated three indicators of inter-rater correspondence. First, codes for article 1 coded by the faculty reviewer and researcher during the pilot study were compared. Codes for article 2 for the same reviewers established inter-rater agreement for one post-training study. Finally, the pilot study codes for article 1 were compared to a post-training review of the same article. For this analysis, correspondence between faculty reviewer, researcher and secondary reviewer were evaluated.

Calculation of inter-rater correspondence was determined by a systematic comparison of coding results. An agreement was defined as both observers selecting the exact same ratings. Reliability was calculated by dividing agreements by agreements plus disagreements and multiplying by 100 percent. Data analysis included a comparison of pre-and post training inter-rater agreement.

Stoiber (2002) reported that procedures for inter-rater correspondence were added to the Manual that would promote reliability for reviews conducted subsequent to Task Force trials. The Procedural and Coding Manual (The Task Force on Evidence-Based Interventions in School Psychology, 2003) was used for coding in this study. Procedures for inter-rater agreement were not found in the appendix of this version of the Manual. Based on this observation, the researcher concluded that these procedures are still under development and may be included with a later version of the Coding Manual.

Consensus Code Agreement

A procedure was conducted to determine a consensus of codes for article 1 and article 2. Following the training and study coding phases of the project, a meeting was held to review the results of each study. The faculty reviewer and researcher re-evaluated each study in comparison to Manual criteria to determine a consensus of "accurate" codes. Previously coded responses by each reviewer were also considered. A second faculty member, Dr. Gary Gottfredson, who served on occasion as an expert consultant also confirmed consensus between these two investigators. Dr. Gottfredson was chosen as a consultant for this project due to his expertise in program development and evaluation research. The results of this review were used as a basis to measure pre-and post-training consensus code agreement for the three reviewers.

Research Review Meetings

Prior to the Workbook training, reviewers met to discuss research procedures and materials. Meeting attended by the faculty reviewer and researcher were conducted to review results of the fall 2003 pilot study and discuss methods for Workbook training and Manual coding. Separate meetings were held by researcher to train the secondary reviewer in these coding procedures.

The faculty reviewer and researcher conducted three meetings to review results of the Workbook and Manual coding. Workbook review sessions focused on discussions about reviewer codes in light of answers provided by the Task

Force. Meetings to review studies coded with the Manual comprised a systemic comparison of ratings and a calculation of inter-rater agreement. As part of this process, researchers discussed their rationale for coding decisions while noting any references to Workbook illustrations. Problems encountered with interpreting coding criteria were documented for reporting of study results. Discrepancies in coding selections were indicated as areas to be addressed in the Coding Workbook. All data, including coding protocols and observation logs were obtained at research meetings.

Data analysis

Data collected during implementation of the Coding Workbook and Procedural and Coding Manual were analyzed to develop a descriptive report. Pre-and Post training quantitative data measured increases in reviewers' ability to apply Manual criteria after Workbook instruction. Qualitative information obtained from reviewer commentary was reviewed to report on Workbook usefulness based on an analysis of its demonstrated strengths and challenges. Data analysis was conducted to examine the degree to which the Workbook increased to reviewer's understanding of Manual criteria. Results of the study were based on analysis of the following data: a) a comparison of pre- and post-training inter-rater correspondence rates, b) a comparison of pre- and post-training coding times and c) a comparison of pre- and post-training consensus code agreement. Reviewer observation logs comprised qualitative information for the descriptive report.

CHAPTER 4

RESULTS

This chapter presents the results of the project undertaken to evaluate *the Coding Workbook for Evidence-Based Interventions* (Shernoff, & Kratochwill, 2003). The results are organized into three sections. The first section will focus on results generated during review of the Coding Workbook. Reviewer's coding accuracy for Workbook exercises and feedback about the conducting the training module will be reviewed. The next section examines use of the Coding Manual prior to and following Workbook training. These results are organized by the following research questions: a) is there an increase in inter-rater correspondence following use of the Coding Workbook; b) is there an increase in consensus code agreement following use of the Workbook training; c) is there a decrease in the time required to code studies following Workbook instruction? The final section focuses on reviewer commentary about the Coding Manual. It investigates reviewer observation logs recorded in pre- and post-training conditions. Trends that emerged in reviewer's commentary throughout the study are presented.

Part I. Review of the Coding Workbook

Results of the Workbook training module are reviewed in this initial section. First, quantitative data including reviewers' coding accuracy is reported. Next, a detailed summary of reviewer commentary is presented. The final section addresses merits and shortcomings of the Workbook.

Coding Accuracy for the Workbook Exercises

Responses to Workbook exercises were compared to pre-determined codings to determine accuracy rates for each reviewer. The total number of responses for the Workbook exercises was 197. Since the Single Participant exercises represented 162, or 82% of those responses, these responses contributed most strongly to the overall accuracy rates. Within this domain, the Key Features criteria comprised the highest number of possible responses or 114 coding selections.

A similar accuracy rate for coding Workbook exercises (88%, 83%, and 84%) was achieved by the three reviewers as presented in Table 2. A comparison of exercises by research design demonstrated that the General Characteristics exercises for Single-Participant Design studies resulted in the lowest overall subcategory accuracy rates for all three reviewers. The researcher and secondary reviewer achieved particularly low accuracy rates of 56 and 44 percent, respectively, when coding Criteria for Other Design Characteristics (i.e. Unit of Assessment) and Statistical Treatment (Familywise Error Rate Controlled). Conversely, reviewers were more successful at determining codes for General Design Characteristics (i.e. Random vs. Nonrandomized design) and Comparison Group exercises. Problems also arose for the coding of Key Features exercises (63%) by the researcher. Within this domain, criteria for Measurement and Statistical Significance of Primary/Secondary Outcomes resulted in the highest number of inaccuracies for this reviewer.

Table 2

Coding Accuracy for the Workbook Exercises

Type of Criteria	**Faculty reviewer**			**Researcher**			**Secondary reviewer**		
Single Participant Designs	N Correct	Total	Percent	N Correct	Total	Percent	N Correct	Total	Percent
General Characteristics	7	9	**78**	5	9	**56**	4	9	**44**
Key Features	102	114	**90**	99	114	**87**	97	114	**85**
Other Descrip.	34	39	**87**	35	39	**90**	36	39	**92**
Total SPD	143	162	**89**	139	162	**86**	137	162	**85**
Group-Based Designs									
General Characteristics	9	11	**82**	9	11	**82**	9	11	**82**
Key Features	21	24	**88**	15	24	**63**	20	24	**83**
Other Descrip.	-	-	**-**	-	-	**-**	-	-	**-**
Total GBD	30	35	**86**	24	35	**69**	29	35	**83**
Total SPD/GBD & W. Average	173	197	**88**	163	197	**83**	166	197	**84**

Reviewers' Observations about the Coding Workbook

This section discusses an overview of the themes that emerged in reviewers' observation logs about Workbook exercises. The discussion will begin with reviewers' feedback about the design of the Coding Workbook. Next, a summary of issues raised about Workbook coding exercises follows these comments. While the discussion emphasizes impressions expressed by all reviewers, it also makes note of points made by a single reviewer that are essential to the evaluation of Workbook usefulness. Coders' understanding of the Manual criteria as well as the ease or difficulty they experienced in applying the criteria is addressed. The review concludes with a summary of strengths and issues documented by reviewers.

Critique of the Coding Workbook Design

Overall, reviewers found that the Workbook was well organized, comprehensive in its scope of Manual criteria and easy to use. The introduction titled *Organization of the Coding Module* provides reviewers with clear and straightforward instructions for coding exercises. The Workbook organization, based on the three strands of coding criteria, (General Characteristics, Key Features and Other Descriptive Criteria) in the Manual, is another one of its strengths. Additionally, the page layout promotes a clear distinction between exercises components and facilitates ease of task completion.

Reviewers found the organization of the coding criteria in the Workbook to be an improvement over the second version of the Procedural and Coding Manual (The Task Force on Evidence-Based Interventions in School Psychology, 2003). In the Manual, overall numerical criteria for Strong, Promising, Weak or No Evidence are presented before the subcategory criteria, while the Workbook locates the general numerical criteria after subordinate ratings. Since the later organization corresponds with the order of the coding decision process, reviewers found this feature to be a considerable improvement to the Procedural and Coding Manual.

The first draft of the Workbook used in this study included documentation of "correct" answers on the Workbook protocol. The protocol was generally effective at providing coders with instructive examples and explanations. The secondary reviewer frequently referred to answers shown on the response protocol to understand criteria requirements and verify coding decisions. Reviewers also reported that explanations for "correct" responses were useful but not available for all Workbook exercises. In addition, the explanations sometimes failed to provide sufficient information for understanding "correct" responses. Although the explanations provided a basis for understanding "correct" answers, they did not supply reviewers with methods used to determine those responses. The Statistical Treatment exercise for Group-Based Designs serves as one example of this concept. Although the Workbook informs readers that the study utilized an insufficient sample size, it does not provide details for interpreting the chart necessary to determine the accurate response.

Reviewers noted other problems with the Coding Workbook protocol. Although designated as a "workbook", the Coding Workbook protocol contained pre-determined responses. This design feature would prevent reviewers from entering responses directly in the Workbook. The protocol was revised to exclude "correct" responses for the present study. The faculty reviewer and secondary reviewer entered responses in the Workbook. However, the researcher followed Workbook directions and entered responses on a separate protocol from the Coding Manual. The procedure of using the Manual instead of the Workbook protocol resulted in a number of complications that included: a) considerable difficulty justifying structural differences between the Workbook and the Manual (Shernoff & Kratochwill, 2003) protocols, b) missed or erroneous coding responses, and c) premature observations of "correct" responses. In addition to these issues, variations in coding instruments required that hand-written revisions be made to the Workbook protocol. These changes resulted in additional coding time and less precise documentation of responses. Overall, using different protocols complicated the task of comparing responses for inter-rater agreement and reviewer accuracy.

Assessment of the Coding Workbook Exercises

Reviewers' feedback about exercises for each category of coding criteria produced mixed results. Some exercises required straightforward coding decisions. However, reviewers encountered difficulty with exercises that included insufficient information for proficient coding.

Reviewers' observation logs indicated that most Other Descriptive Criteria exercises were "clear and straightforward". Criteria labeled as "easy to code" in this report was easily interpreted and applied to study criteria by reviewers. For this category, the clarity of study excerpts and Manual criteria promoted clear-cut coding decisions. Feedback reflecting reviewers' confidence in coding decisions was consistent throughout their observation logs. Criteria required further clarification for only a few of the exercises. Reviewers' feedback about Other Descriptive Criteria exercises is presented in Appendix A.

Overall, the Workbook was helpful in extending reviewers' understanding of the coding criteria. The combination of exercises and "correct" responses in the Workbook was described as its most effective component. In many cases, self-instruction was promoted through associations between coding criteria and the "correct" responses. Explanations for these responses increased the effectiveness of the exercises, particularly when reviewers had prior knowledge of the research method under review. Some examples of exercises that illustrate this point include Statistical Treatment, Measurement (Group-Based Designs), and Quality of Baseline. The selection of study excerpts was also beneficial to the training process. Studies that appear in the Workbook are generally straightforward, which is an aspect that enables reviewers to focus on interpreting criteria and to code with greater accuracy.

Reviewer commentary also noted issues that arose during Workbook training. For example, the researcher and secondary reviewer suggested that coding many of the General Characteristics and Key Features exercises required a higher level of expertise in research methodology than was possessed by either graduate student. The researchers encountered the most difficulty with coding Other Design Characteristics (coding assignment of participants to conditions of a Single-Participant Design study) Statistical Treatment, Measurement (Reliability and Validity) and Statistical Significance. For Statistical Treatment exercises, researchers were not able to code criteria for Familywise Error or Sufficiently Large N with information provided in the Workbook. Descriptions of Cohen's conventions for effect size and methods for calculating sufficiently large N were not available to reviewers in the Manual or in the Workbook. A summary of reviewers' commentary regarding General Characteristics exercises can be found in Appendix B.

Considerable problems arose for Measurement criteria. For example, the Measurement exercise criteria for Reliability states “The evidence for psychometric properties must be reported or referenced in the article” (The Task Force for Evidence-Based Interventions, 2003). However, it also states that “Observable incidence and/or occurrence rates, such as school attendance rates, homework completion rates or other *well-known* [italics added] standardized, norm-referenced assessments will be considered reliable measures” (The Task Force for Evidence-Based Interventions, 2003). To code these exercises, reviewers must have prior knowledge of measures considered “well-known” by the field. If Reliability is not reported by authors, then coders must either make subjective coding decisions or conduct the time-consuming process of referencing previous research. Similar problems arose for Validity exercises. Coding criteria in this area requires knowledge the empirical and theoretical basis of assessments used in the study. Other problems encountered with this and other Key Features exercises are reviewed in Appendix C.

Commentary by the researcher and secondary reviewer indicated that additional information about statistical calculations must be available in the Workbook to promote competent coding by Masters level graduate students. As part of evaluating Key Outcomes Significant for Group-Based Designs, reviewers must chart information for a variety of indicators including Reliability, Effect Size (ES), and 1-β. The criteria descriptions provide little guidance for calculating factors that are reported in the study. Reviewers are referred to the Manual for guidance in the calculation of Effect Size and appropriate sample size. However, the researcher and secondary reviewer (both graduate students) stated that they lacked the research background necessary to interpret the statistical formulas contained in these charts. In order to code this exercise, reviewers had to consult the answer key. Overall, reviewers' opportunities to learn through first interpreting coding criteria and following this process with response verification were greatly reduced.

The need for enhanced coding instructions, criteria clarifications, and response key information was pervasive throughout reviewer comments. The graduate student reviewers who had limited experience with evaluating scientific research requested clarifications most often. However, the faculty reviewer also identified selected coding criteria that required further explanation. For example, the first Workbook exercise required coders to identify the research design used for a Single-Participant Design study. In this instance, a definition for "Simple vs. Complex Phase Change", essential for understanding the criteria, was not presented in the exercise.

Researchers also identified a number of Workbook exercises that failed to provide adequate definitions of research terms. Responses to these exercises were frequently inaccurate due to reviewers' multiple interpretations of coding criteria. Difficulties encountered by the researcher with applying Implementation Fidelity criteria for Group-Based Designs exercise is one example of this concept. The Manual criteria maintain that evidence of Acceptable Adherence must be measured through ongoing consultation/supervision, coding sessions or audio/video taping. The researcher, based on graduate school training, interpreted the meaning of consultation/supervision as “face to face” collaboration. However, the criteria refer more specifically to written documentation, a standard that was satisfied by measures used in the study. Consequently, the alternate interpretation by the researcher resulted in a coding error.

Reviewers noted differences in criteria descriptions that impacted Workbook coding. The researcher was able to identify a number of exercises comprising criteria descriptions that were less comprehensive that corresponding criteria presented in the Manual. Manual criteria that comprise charts or explanations considered informative to Workbook coding include Validity, Educational Significance and Identifiable Components. Refer to Appendix D for a complete listing of editorial suggestions for all Workbook exercises.

Although most Other Descriptive Criteria exercises were straightforward, this category did result in some difficulties. The Intervention Style exercise was cited as one example. This exercise requires reviewers to identify theoretical orientations for intervention programs. Although the Workbook lists theoretical orientations, it fails to provide related descriptions. The straightforward nature of the study methods enabled all three reviewers to code the exercise accurately. However, reviewers also noted that coding other school-based studies might require a listing of theoretical orientations. Appendix A comprises a summary of additional reviewer feedback about Other Descriptive Criteria exercises.

With regard to gaining proficiency with coding complex or “difficult to code criteria”, reviewers noted the Workbook failed to provide sufficient practice opportunities for evaluators who lacked expertise in research methods. For this report, “difficult to code” criteria are measures that resulted in consistent coding errors or notations of problems by reviewers. As one instance, the graduate student researchers reported experiencing considerable difficulty with the Statistical Treatment exercise. However, the Workbook only includes one such exercise. In addition, criteria for the Type and Stage of Program are not addressed in the Workbook. Coding many school-based studies is likely to require a high level of reviewer competency. To that end, reviewers stated that additional practice with coding more challenging exercises is needed for proficiency in reviewing school-based research.

Summary of Reviewers’ Overall Impressions of the Workbook

Reviewers’ observation logs revealed a number of themes regarding their thoughts about the Workbook training exercises. The coding task was assessed as straightforward for some exercises but more difficult for others. Specifically, coding was easiest on exercises with clearly written study excerpts, well-developed criteria and informative response protocols.

Accuracy data and reviewers’ observation logs indicated multiple areas of difficulty. The researcher and secondary reviewer often attributed these difficulties to a limited knowledge of research methodology. However, they also stated that information to extend their understanding of complex Manual criteria was sometimes insufficient. As such, the secondary reviewer found that inspecting the response key was often necessary to understand Workbook exercises. Although “correct” answers were provided for all the exercises, only a few exercises also provided explanations for responses that were instructive to reviewers. These illustrations were most critical for calculating statistical outcomes and other more complex study methods.

All reviewers experienced problems with the absence of term definitions, insufficient criteria explanations and unclear coding directions. Other Descriptive Criteria exercises that were otherwise characterized as "easy to code" sometimes lacked necessary term definitions for Workbook criteria. In addition, inconsistencies between coding criteria and accurate responses indicated by the Task Force resulted in varying degrees of confusion for reviewers.

Part II. The Pre- and Post-Training Review with the Coding Manual

Research data obtained from Manual coding conducted prior to and following Workbook training is reviewed in this section. These results are organized by the following research questions: a) is there an increase in inter-rater correspondence following use of the Coding Workbook, b) is there an increase in consensus code agreement following use of the Workbook module, c) is there a decrease in the time required to code studies following Workbook instruction? These results comprise quantitative data annualized to assess usefulness of the Coding Workbook.

1. *Is there an increase in inter-rater correspondence following implementation of the Coding Workbook?*

As discussed earlier in the Methods chapter, the present study consisted of a pre-training article coding, a Workbook training exercise and a post-training article coding. The pilot study comprised coding of article 1 (Larson, 1989) by the faculty reviewer and researcher. Then, the Workbook training was conducted by these reviewers and by the secondary reviewer. Following Workbook training, article 2 (Seifer et al., 2004) was reviewed by the faculty reviewer and researcher. Article 1 was coded as the post-training measure for the secondary reviewer.

Results of the Pilot Study

Inter-rater reliability for the pilot study was determined during a research review meeting attended by the faculty reviewer and researcher. A comparison of coding decisions by each reviewer yielded an inter-rater agreement of 59 percent. In total, reviewers selected identical responses for 109 of the 186 possible criteria. Rates of inter-rater correspondence for three categories of Manual criteria are presented in Table 3. In terms of specific criteria, disagreements for the pilot coding were highest in the areas of Measurement, Length and Intensity/dosage of Intervention, and Training/Support Resources.

The review meeting prompted discussion that enabled the reviewers to develop a better understanding of the Manual criteria. Discrepancies in coding decisions lead to closer analysis of the criteria and exchange of information between reviewers. As such, consultation between reviewers was effective for clarifying criteria and providing knowledge of research methods that was not available during independent coding.

Results of the Post-Training Study

Inter-rater agreement was calculated for two post-training measures. The coding of article 2 by the faculty reviewer and researcher is considered initially. This post-training review yielded an inter-rater correspondence that was comparable to the rate established during the pilot study. The total number of possible codings varied from the pilot study due to differences in study design. Where the pilot study produced a total of 186 responses, the second coding study generated 165 total responses. For the post-training review, evaluators selected identical responses for 98 of 165 criteria. Thus, the inter-rater reliability reached 60 percent for the post-training measure versus 59 percent for the pilot study. Table 3 presents a summary of the inter-rater correspondence for this phase of the study.

Table 3

Inter-rater Agreement Percentages for Faculty Reviewer and Researcher (Manual Coding)

Type of Criteria	Article #1 (Pre-Training)	Article #2 (Post-Training)
General Characteristics	73%	64%
Key Features	61%	59%
Other Descriptive Criteria	52%	60%
Total	59%	60%

Inter-rater agreement rates were also generated for the post-training review of article 1. The responses by the secondary reviewer were compared to those selected by the faculty reviewer and researcher to arrive at corresponding figures. These results presented in Table 4 indicate low rates of inter-rater agreement.

Table 4

Inter-rater Agreement Percentages for the Pre- and Post-Training Reviews

Type of Criteria reviewer	Secondary reviewer vs. Faculty reviewer	Secondary vs. Researcher
General Characteristics	55	64
Key Features	47	50
Other Descriptive Criteria	59	68
Total	50	56

2. *Is there an increase in coding consensus agreement following use of the Workbook module?*

Reviewer correspondence to “consensus” responses was also examined. Codes for both articles were determined by a consensus of the reviewers and a third faculty member. These rating were then compared to each reviewer’s responses from Manual coding. Results for pre-and post-training coding for all reviewers are presented in Table 5.

Table 5

A Comparison of Reviewer Consensus Code Agreement Percentages for Manual Coding (Pre-and Post Workbook Training)

Manual Reviewer	Faculty Reviewer		Researcher		Secondary
Section (Post WB)	Article 1	Article 2	Article 1	Article 2	Article 1
General Characteristics	*91*	91	*82*	73	*73*
Key Features	*86*	75	*80*	75	*55*
Other Descript.	*84*	100	*75*	70	*68*
Total	*86*	82	*79*	74	*59*

A comparison of pre- and post-training consensus code agreement for article 1 provides useful information about effectiveness of the Coding Workbook. Overall, reviewers' agreement with consensus code responses did not increase after completing of the Workbook training. A comparison of results for pre- and post-training review of article 1 suggests that consensus code agreement decreased following use of the Coding Workbook. This measure for the secondary reviewer was 27 percentage points less than the faculty reviewer and 20 percentage points less than the researcher.

The General Characteristics criteria resulted in the highest consensus code agreement for all three reviewers. The largest difference in this measure by the secondary reviewer was calculated for Key Features criteria. Within this area, disagreements were highest for Measurement and Primary/Secondary Outcomes are Statistically Significant criteria. Observation logs provided insight into the difficulties encountered by this reviewer. Coding errors in this area appeared to result from misinterpretations of criteria and a failure to fully document study information.

Some interesting patterns emerged within criteria categories coded by the faculty reviewer and researcher. The agreement to consensus codes for General Characteristics criteria by the faculty reviewer reached a satisfactory rate of 91 percent for both articles. This reviewer's consensus code agreement for Other Descriptive Criteria increased to 100 percent during the second coding. These results indicate a competent coding ability in these areas. Conversely, the researcher achieved lower agreement to consensus codes for both criteria.

3. *Is there a decrease in the time required to code studies following Workbook instruction?*

All reviewers documented coding times in observation logs. Coding time for each review session was added to arrive at a cumulative coding time for each article.   Table 6 lists coding trial times for each reviewer.

Table 6

Time (Hours) Required for Coding Research

Research Study	Faculty Reviewer	Researcher	Secondary Reviewer
Article 1	4.5	6.75	7.5
Article 2	2.6	5.0	-
Article 3	-	3.0	-
Article 4	-	5.25	-
Average	4.0	5.0	7.5

The coding task proved to be a challenging task for all researchers. During the pilot study, coding of article 1 took 4.5 hours for the faculty reviewer and 6.75 hours for the researcher. The recorded time for each reviewer includes 30 minutes for reading of the research article. The secondary reviewer coded the same article following the self-training module. The total time for this review was 7.5 hours. The post- training time required for coding by the secondary reviewer was comparable to the pre-training coding time for the researcher. As such, coding practice obtained during the Workbook training did not appear to effect subsequent coding times for this reviewer.

Post-training study review by the faculty reviewer and researcher did result in reduced coding times. The faculty reviewer completed coding of Article 2 in 2.6 hours, which was approximately 2 hours less than the time it took to code article 1. The researcher who coded the same article in 5 hours achieved a comparable reduction in coding time. Since considerable time was needed to record detailed commentary, the researcher documented this time separately. The total time dedicated to this effort amounted to 1 hour and 15 minutes. Decreases in coding times documented in this study are consistent with the time reductions generated as a result of practice effects described by Stoiber (2002). However, the shortest coding time of 2.6 hours in this study was still longer than the coding time of 2 hours recorded by the Task Force (Stoiber, 2002).

Reviews of subsequent articles by the researcher resulted in even greater reductions in time. Article 3, a PATHS research study conducted by Kam, Greenburg and Walls (2003) was read and coded by the researcher in 3 hours. An additional 20 minutes was used to document observations. The two hours required to code this study was a 40 percent reduction in time compared to the time for coding article 2. However, there was high variability in coding times of these studies. The researcher coded article 4, a PATHS study that focused on increasing children's emotion competency by Greenberg at el. (1995) in 5.25 hours. This coding time was comparable to that of article 2. Recoding observations was completed in an additional 30 minutes. Overall, reviewers did reduce coding times following their initial study review.

Part III. Reviewer Commentary about the Coding Manual

Summary of Reviewers' Observations for the Pilot Study

Review of article 1 (Larson, 1989) was conducted by the faculty reviewer and researcher. The observation logs recorded during this review comprise the reviewers' initial impressions of the coding with the Manual. In their initial commentary, reviewers described the Manual as impressive but also overwhelming and intimidating.

Reviewers' commentary provides insight about each coder's ability to interpret Manual criteria. The observation logs of both reviewers indicated areas requiring further clarification, clearer instructions and explanations. As will be evident throughout study, the varied experience levels of reviewers resulted in differences in commentary content. Where comments by the faculty reviewer illustrated sufficient background knowledge for Manual coding, the researcher's observation logs were largely focused difficulties with interpreting coding criteria.

Despite the difficulties reflected in reviewers' observation logs, the consensus code data indicates that reviewers were fairly successful at interpreting the Manual criteria. However, this data also show that coding was a greater challenge for the researcher than it was for the faculty reviewer. The overall agreement with consensus codes for the faculty reviewer reached 86 percent while the same figure for the researcher was calculated at 79 percent. Both reviewers achieved their highest consensus code agreement for General Characteristics criteria. Conversely, the lowest rates were recorded for Other Descriptive or Supplemental Criteria.

While some aspects of the Procedural and Coding Manual could be readily interpreted by the graduate student, some criteria posed significant problems. The researcher described the coding trial as “the first opportunity to apply concepts learned in statistics” and a “very challenging task”. Difficulties were encountered with applying coding considerations in the General Characteristics and Key Features sections of the Manual. Specifically, questions arose when applying criteria for statistical data analysis and measurement of primary/secondary outcomes. Issues were also raised with regard to Statistical Treatment (Unit of Analysis and Sufficiently Large N), and Measurement (Reliability and Validity) criteria. Despite efforts that included reference of class notes and charts in the Manual, the researcher had limited success with coding these criteria.

When research methodology was not clearly defined in the study, other problems emerged that potentially reduced the coding validity. The graduate student conducting the case study found that subjective interpretations were required for information that was omitted or not explicitly stated in the research. A number of these problems arose during the evaluation of Key Features. In these cases, selections were often based on the evaluator’s “best judgment” or decided by electing criteria as “unknown” for missing or unclear information. Levin (2002) has found that indistinct coding criteria may result in decrements to inter-rater agreement.

Issues raised by the faculty reviewer focused on similar Manual criteria but varied in content. Commentary by this reviewer centered on more technical questions than explanations for approaching criteria coding. Concerns arose with regard to necessary clarifications of Manual criteria that ranged from the absence of term definitions to more problematic issues of vague coding instructions.

Summary of Reviewers' Post-Training Observations

Reviewers' commentary about the post-training review was recorded in observation logs. The information that was analyzed and compared to coding consensus results provides further insight about the usefulness of the Workbook training module.

Assessment of the Post-Training Review by the Secondary reviewer

Following the Workbook Training, article 1 was coded by the secondary reviewer. Commentary by this reviewer indicates difficulties with coding this article for a variety of Manual criteria. Consensus code results support this finding. The secondary reviewer coded article 1 with a consensus code agreement rate of 59 percent, which is a substantial decrease from the Workbook coding accuracy by this reviewer of 84 percent.

Comments by the secondary reviewer emphasized a general lack of confidence in her knowledge of and ability to code using the Manual. The reviewer stated that the trial served to identify a number of areas for which additional instruction and training were necessary. With regard to specific training, the reviewer has taken 2 statistics courses but emphasized that coding with the Manual may require a review of those courses and additional instruction through a third statistics course.

Criteria identified by the secondary researcher as most “difficult to code” were similar to those described by the researcher. These criteria include: Statistical Treatment, Type of Program, Measurement and Primary/Secondary Outcomes are Statistically Significant. Difficulties with using Table 1 in the Manual accounted for problems with Statistical Treatment ratings. Problems with identifying primary and secondary measures and coding corresponding criteria were similarly noted in commentary of all reviewers. The secondary reviewer stated that the charts were “time consuming” and “difficult to fill out” in the absence of a completed sample. This reviewer also had difficulty interpreting outcome statistics described in the article 1.

Assessment of the Post-Training Review by the Faculty Reviewer and Researcher

Commentary by the faculty reviewer and researcher document their observations about coding article 2. The criteria indicated “easy to code” by reviewers remained largely similar to the pilot study. The researcher noted having a clearer understanding of some research methods that were unfamiliar during the initial study. Much of this knowledge was obtained through research meetings with the faculty reviewer. Despite these improvements, reviewers’ difficulties with coding article 2 remained largely unchanged from those indicated for article 1. As such, reviewers did not develop the ability to code “difficult” criteria as a result of the Workbook training.

Strengths of the Coding Manual

Commentary by reviewers identified a number of criteria as “clear” and “easy to code”. These criteria were described by reviewers as clearly defined in the Manual and easy to apply to research examined in this study. The General Characteristics criteria that were consistently described as “clear and understandable” include General Design Characteristics and Historical Intervention Exposure. Within Key Features criteria, only the Site of Implementation received this distinction. A number of criteria in the Other Descriptive area that were similarly identified include: Length of Intervention, Dosage Response, Characteristics of the Implementer, Cost Analysis Data and Feasibility.

Differences in reviewers’ perceptions of straightforward criteria were found for the Other Descriptive Criteria category. The faculty reviewer described most criteria in this area as straightforward. However, problems with coding External Validity Indicators and Intervention Orientation criteria were noted by the researcher.

Observations recorded by the researcher emphasized the Workbook exercises as an asset to the review process. In particular, the Workbook provided a reference for coding similar criteria in the intervention under study. The exercises, which utilized straightforward study excerpts, served as a clear example of criteria applications. References to the Workbook exercises facilitated the coding of General Design Characteristics, Statistical Treatment and Measurement criteria. In a broader sense, these comments suggest that school psychologists can use the Workbook as a tool for coding studies following completion of the training module.

Issues Indicated for the Manual Criteria

Statistical Treatment criteria were indicated by both coding trials as an area of difficulty for the researcher. The criterion subcategories of Appropriate Unit of Analysis and Familywise Error Controlled were both coded inaccurately. For both criteria, the information needed for coding was not clearly stated in the article. Appropriate Unit of Analysis assesses the level of program implementation in the study in comparison to the intervention model. In the article under reviewer, authors characterized the PATHS program as "typically" implemented on a universal level. However, other studies, one conducted by program developers, have implemented PATHS as a selective intervention (Greenberg, et al., 1995). Therefore, the researcher assumed that the selective implementation of the program described in the study could be considered as appropriate. This assumption was later determined to be inaccurate during the second review meeting. This outcome illustrates the potential for multiple interpretations of criteria that can result in the absence of clear coding instructions in the Manual.

Coding decisions with regard to Statistical Significance of Primary and Secondary Outcomes were consistent areas of difficulty for reviewers. Determination of this these criteria requires the completion of two charts. To complete these charts, reviewers must accurately distinguish between primary and secondary outcomes. Although the Manual provides guidelines for this procedure, it proved to be repeatedly difficult for reviewers. Additionally, each chart requires the recording of various indicators such as Reliability, Effect Size and Sufficiently Large N or sample size. Interpretations by both reviewers led to discrepancies with consensus codes in the classification of Primary and Secondary measures. In turn, these miscodes led to further problems in the coding of Statistically Significance of Primary/Secondary Outcomes. Because Statistically Significant Outcomes comprise a major part of the Key Features criteria, miscodes by reviewers had a significant impact on the overall outcomes for this category. Based on similar results obtained during the pilot study, the reviewers concluded that independent coding of these criteria was difficult for both experienced and novice reviewers.

Identification of Primary and Secondary measures also impact coding decisions about Measurement. Reviewers miscoded criteria for Validity as well as the general rating for Measurement. Both of these criteria are based on decisions about Primary measures. The consistency of these difficulties indicates that the Workbook exercises failed to provide adequate instruction to reviewers in this area.

Reviewers encountered problems with Implementation Fidelity criteria. Problems with coding these criteria occurred for both Manual reviews and Workbook training. Coding errors, as indicated by reviewer commentary and the researcher's review of the Manual, were related to imprecise criteria descriptions and coding instructions. Variations in coding instructions for Acceptable Adherence led to differences in reviewers' interpretation of these criteria. Coding instructions for Implementation Fidelity in the Manual state "to receive a rating of 3 or strong evidence, the study must demonstrate acceptable adherence. *In addition*, [italics added] evidence should be measured through at least two of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes, *and* use of a manual." However, criteria specific for Acceptable Adherence seem to contradict this standard by stating that Acceptable Adherence is met through use of the procedures above. While reviewing the article, reviewers focused on different aspects of the criteria and therefore selected dissimilar coding responses.

Reviewer commentary also illustrated difficulty with applying Manualization criteria to article 2. The Manual states "The candidate intervention should be "manualized" (i.e., accompanied by a clear description of the procedures used) and the studies must be conducted with intervention manuals and/or detailed procedural specification" (Kratochwill &Stiober, 2003). PATHS is a manualized intervention and authors do list a reference for the PATHS curriculum in the candidate study. However, the article only documents use of training workshops and teacher materials. In addition, ongoing supervision by a PATHS consultant was conducted. This discrepancy between the implication of manual use made by the PATHS curriculum reference and the more informal study description resulted in different coding decisions by the reviewers.

Other problems arose with Manual criteria that required reference to previous research on the intervention under review. As part of evaluating the Stage of Program and Replication criteria, reviewers must classify the developmental stage of the study and the degree to which it is a replication of prior research. For the present trial, neither reviewer could find adequate documentation of prior research in the study. Problems with applying these Manual criteria to article 1, which included limited documentation of previous research, were also noted by the secondary reviewer. Reviewers must then decide whether it is appropriate to search the literature for prior studies of the intervention. References to Manual criteria failed to yield a resolution for this problem. Coding decisions for these criteria included a literature search by the faculty reviewer but not the researcher. Consequently, low rates of inter-rater agreement were recorded for this criterion. In relation to this concern, reviewers' emphasized the need for guidelines to address this issue and similarly vague descriptions of previous research. Appendix E presents other issues related to coding with the Manual.

CHAPTER 5

DISCUSSION

The present case study provided information about the effectiveness of the Workbook as a training instrument for the EBI coding process. In addition to observations of the process, measures of inter-rater correspondence and reviewer consensus code agreement collected prior to and following implementation of the Workbook were used to evaluate its utility. The study failed to produce post-training improvements to inter-rater correspondence or consensus code agreement. Increases in coding performance were not demonstrated for the pre- and post-training review of article 1 or the post-training review of article 2. The case study results did indicate reductions to coding time for articles reviewed by two of the researchers. Finally, reviewer commentary generated considerable information about the issues related to coding with the Manual and Workbook.

This chapter will discuss results of the case study in an attempt to assess the current usefulness of the Coding Workbook. It will begin with a discussion of results from the Workbook training. Next, quantitative data from the pre-and post-training reviews is considered. This section is organized by research questions: a) is there an increase in inter-rater correspondence following implementation of the Coding Workbook, b) is there an increase in reviewers' consensus code agreement after implementing the Workbook module, c) is there a decrease in the time required to code studies following Workbook instruction? Then, the discussion will focus on implications of reviewers' commentary about the Workbook and Manual coding. Finally, an examination of the difficulties that emerged during the coding process will be used to formulate suggestions for improvements to the Coding Workbook.

Evaluation of Results for the Coding Workbook

In general, coding of Workbook exercises by the three reviewers yielded similar percentages of accuracy. Coding accuracy by the researcher and secondary reviewer differed by only one percentage point. This is a surprising result given the difference in coding experience between the two researchers. Previous experience by the researcher comprised using the Manual to code article 1 and reviewing these results with the faculty reviewer. Conversely, the Workbook module served as the secondary reviewer's first exposure to the Manual criteria.

With regard to type of Manual criteria illustrated by Workbook exercises, inaccuracies were highest for General Characteristics in Single-Participant Designs (78%, 56% and 44%) and for Key Features in Group-Based Designs (63%). The number of Single-Participant Design exercises in General Characteristics accounts for this low accuracy percentage. Reviewers had considerable difficulty with the exercise for "Unit of Assignment to Conditions", which comprised 7 of the 9 possible coding decisions. The low accuracy percentage for Key Features exercises coded by the researcher can be attributed to difficulties with the Primary/Secondary Statistically Significant Outcomes exercise. The researcher lacked the background in statistical methods necessary to understand these criteria.

Reviewer commentary demonstrated that the graduate students lacked the prerequisite background knowledge for Workbook coding. As such, a higher level of training and experience is needed for reviewers to understand and benefit from the training module. A discrepancy between the level of research experience required for coding and the background of these reviewers is one possible reason for this problem. The Workbook was initially designed for training of school psychologists with a background in research methodology. As graduate students who are currently at the Specialist level of training, the researcher and secondary reviewer have obtained limited research experience obtained through coursework and graduate assistantships. Although the graduate students have taken two statistical courses, they have had limited opportunities to apply this knowledge.

The faculty reviewer demonstrated an understanding of the Workbook criteria that was superior to that of the graduate students. In addition to doctoral level training, the faculty reviewer has substantial research experience. As a result, the observation logs by this reviewer illustrate a proficient understanding of the research methods applicable to the Workbook criteria and an ability to evaluate the usefulness of the coding criteria. Approaching the task from this perspective, the faculty member still encountered difficulty with a completing a number of Workbook exercises. The need for enhanced coding instructions, criteria explanations and term definitions were emphasized in feedback provided by this reviewer.

Analysis of the Pre- and Post-Training Study Coding

Was there an increase in inter-rater correspondence following implementation of the Coding Workbook? Case study results for pre-training review of article 1 and the post-training review of article 2 demonstrate consistent low rates of inter-rater agreement. The inter-rater agreement of 60 percent reached during the post-training measure is similar to the rate of 59 percent achieved during the pilot study. In addition, the post-training review of article 1 yielded similar low rates of inter-rater agreement. These figures are significantly lower than the concordance rates for Manual coding documented by the Task Force. Reviews conducted by Task Force members and graduate students produced inter-rater agreement rates averaging 85 percent.

There are several possible reasons for the low rates of agreement in this study. One reason is the time period between training and post-training study review. The post-training article was coded 1 month after Workbook training by the faculty reviewer and researcher. The secondary reviewer coded article 1 two weeks after training. The length of time to conduct training should also be considered. The procedures used in this study allowed for training schedules to be individually determined by reviewers. As such, the time required for reviewers to complete Workbook exercises varied from two days to one month. During the time period from the beginning of training to article coding, reviewers could have forgotten or become unclear about knowledge obtained during Workbook training. Outcomes of this study suggest that future training procedures should specify time requirements.

Another potential reason for the low rates of inter-rater agreement relates to the effectiveness of the Workbook training module. It is possible that Workbook training failed to produce an increase in reviewers' knowledge of Manual criteria. The information in observation logs demonstrated that reviewers who were frequently unclear about the coding criteria based their decision on inferential judgment about these standards. Specifically, reviewers' cited these difficulties with coding Measurement and Implementation Fidelity criteria. This general finding is consistent with studies of the Manual conducted by Levin (2002). In these examinations, Levin (2002) observed that inferences about missing or vaguely stated criteria resulted in low rates of inter-rater correspondence.

Differences in inter-rater correspondence are explained in part by variability in consensus code agreement rates. The faculty member achieved a post-training consensus code agreement rate for General Characteristics and Key Features criteria of 91 and 100 percent. Conversely, miscodes by the researcher or secondary reviewer resulted in low rates of inter-rater agreement for these criteria. Hence, low consensus agreement resulted in problems with inter-rater reliability. These results also indicate that additional training is required for the latter two reviewers.

Coding of the third category, Key Features criteria, resulted in a similar consensus code agreement rate for the faculty reviewer and researcher. However, this category yielded an inter-rater agreement of only 59 percent. Although reviewers made errors that were different from each other, these disagreements occurred for specific Manual criteria. Disagreements were highest for Replication (100 percent), Implementation Fidelity (34 percent), Comparison Group (33 percent), and Statistically Significant Primary/Secondary Outcomes (30 percent). These results suggest that the Workbook training failed to produce a clear understanding of specific Key Features criteria for either reviewer.

Was there an increase in reviewers' consensus code agreement after implementation of the Workbook module? The results of post-training coding trials were disappointing. The post-training review of article 1 by the secondary reviewer yielded an overall consensus code agreement of 59 percent. This percentage was considerably lower than the pre-training consensus rates of 86 and 79 percent for the same article. The practice acquired by the faculty reviewer and researcher during the pilot study is one possible explanation for this difference. By participating in the pilot study, these reviewers had opportunities to review Manual criteria and practice coding that was not available to the secondary reviewer. The faculty reviewer and researcher also participated in other activities that utilized review of the Coding Manual.

A second related explanation relates to access to the Workbook responses during Manual coding. Where the researcher referred to Workbook exercises to assist with coding article 2, the secondary reviewer did not have access to the

Workbook responses while coding article 1. Therefore, the reviewer did not have same opportunity as other reviewers to reference coded exercises as guidance for study coding.

Third, the duration of time between training and study coding, as mentioned above, may have resulted in reduced proficiency in coding for all reviewers. Variability in the level of difficulty between Workbook study excerpts and research studies in the present study may also account for low agreement with consensus codes. There were several differences between the Workbook and the study articles that contributed to difficulties with article coding. The Workbook exercises contained study excerpts that were directly applicable to the coding task. Conversely, reviewers coding complete research studies have to pinpoint relevant study methods. This task may be challenging when applied to school-based research that lack clarity in reporting of study methods. The research methods in Workbook study excerpts were usually straightforward. In contrast, the research studies coded for this project were complex in that they utilized a variety of outcome measures for which judgments about classification of primary and secondary measures required a great degree of interpretation by the reviewer. Evaluation of reliability and validity was similarly challenging. These challenges likely contributed to low agreement to consensus codes for article 1 and article 2.

A final reason for disappointing results for post-training results may also be related to effects of the Workbook training. The results demonstrate that none of the reviewers reached a satisfactory level of consensus code agreement after using the Coding Workbook. Workbook training resulted in increased proficiency for only one category of criteria. This suggests that the Workbook may not have been effective in increasing the reviewers' understanding of the Manual criteria. The low rates of consensus code agreement indicate that additional training and coding practice are needed by all reviewers, particularly the graduate students.

Other researchers have described similar difficulties coding Manual criteria. According to Christenson, Carlson and Valdez (2002), accurate coding with the Manual requires considerable knowledge of the literature and competency in research methodology. These authors stated that advanced graduate students and professionals trained in a scientist-practitioner oriented program may be best suited for this task. In a study by Prevatt and Kelly (2004), faculty members had significant difficulties with interpreting Manual criteria and Master's level graduate students lacked the background necessary for coding. Based on these outcomes, the researchers speculated that coding with the Manual may be difficult for school psychologists at all levels of training. The results of this study provide support for this idea. Thus, Workbook may need to be modified to increase its utility for training school psychologists and developing their understanding of the coding criteria.

Was there a decrease in the time required to code studies following Workbook instruction? A shorter coding time was recorded for reviews conducted after Workbook training. Article 2 was coded by reviewers in 2.6 and 5.0 hours, which for both reviewers was approximately 2 hours less than it took to code article 1. Reductions in coding time ranged from 1 hour and 15 minutes to 3 hours and 45 minutes for subsequent study reviews by the researcher. These time reductions suggest that coding studies with the Manual over repeated sessions can result in practice effects and reduced coding times. The average coding times observed in this study are comparable to times recorded by Prevatt and Kelly (2004). However, they fail to approach coding times of 2 hours documented by the Task Force (Stoiber, 2002), which may reflect greater degrees of coding practice than was obtained in the present study.

Even as coding efficiency increases, reviewers may experience variations in coding time related to the complexity of the coding task. In the present study, coding times were impacted by the complexity of study methods and the degree to which information about the study was clearly and completely reported. For example, article 3 is comprised of clearly designed and documented study methods that facilitated efficient coding. In this study of PATHS implementation quality, conducted by the program developers (Kam, Greenberg & Walls, 2003), study characteristics such as research design, outcomes measures, statistical methods and results were explicitly stated in the article. Conversely, a study of the PATHS program (Greenberg et al. 1995) in its early stages of development was coded as article 4. This study, which examined program effectiveness with regular and special education populations, utilized a more complex research design and combination of outcome measures than was used in article 3. Therefore, review of this article was more challenging and time consuming.

Reviewer Commentary about the Coding Manual

Reviewers' commentary about the effects of training on coding studies with the Manual yields important information about the usefulness of the Workbook. Observation logs demonstrated that the Workbook training resulted in minimal changes to reviewers' perceptions of their ability to understand and code Manual criteria. Coding criteria described as "easy" or "difficult to code" remained constant for the pre- and post-training coding trails. In comments that followed Workbook training and subsequent study coding, the researcher and secondary reviewer emphasized a general "lack of confidence" in their ability to interpret difficult Manual criteria. In general, the feedback from all reviewers indicates that the Workbook training did not increase their coding proficiency with the Procedural and Coding Manual.

Areas of difficulty with study coding reported by reviewers have implications for expanding the instructive nature of many Workbook exercises. A number of Workbook exercises were noted by the researcher and secondary reviewer as containing statistical terms or research methods that were advanced for their current level of knowledge and experience. Observation logs recorded by these reviewers emphasized competent coding with the Manual would require additional instructive support and training. To function as a training instrument that is instructive to school psychologists at a similar level of training, the Workbook will need to include definitions of statistical terms and more detailed directions for applying statistical formulas. Explanations of statistical charts and procedures may require the Workbook to present “worked-through” examples as models for reviewers to follow.

Other Workbook exercises resulted in coding difficulties for reviewers regardless of research expertise. Reviewers stated that judgments about criteria such as Appropriate Unit of Analysis or Stage of Program were often complicated by insufficient descriptions of study methods. The absence of specific instructions in the Workbook to addresses these issues led to reductions in inter-rater reliability due to subjective decision making. Coding instructions in the Workbook and Manual should be expanded to aid reviewers’ efforts in evaluating study methods. Similar problems arose with evaluating Validity and Reliability of study measures. According to reviewers, the Workbook must provide more explicit guidelines for evaluating assessments that are not supported by evidence of Validity and Reliability in studies. Furthermore, additional information is needed to aid reviewers’ determination of reliable and valid “well-known” norm-referenced instruments.

Reviewers also had considerable difficulty applying criteria for primary and secondary measures to research study methods. These problems arose when reviewers attempted to distinguish between direct and indirect treatment outcomes. Significant variations in coding decisions resulted from differences in criteria interpretations. To enable reviewers to make these distinctions, the Workbook exercises must provide reviewers with comprehensive explanations of criteria and examples of these concepts. Implementing these reviewer suggestions would facilitate coding decisions and reduce inaccuracies.

One approach to evaluating the Workbook is to define its target consumer. Any revisions to the Workbook should be made in light of the trainees' knowledge and background in research methodology. If designed for training school psychologists with research expertise in Manual coding, a number of clarifications are needed. If the intention is to train graduate students and practicing school psychologists, the Workbook must be supplemented with instructional text that is tailored to the unique needs of these consumers.

Recommendations for Improvements to the Coding Workbook

Suggestions for improvements to the Coding Workbook were developed based on the results of the present case study. Information generated from reviewer commentary and quantitative results of pre-and post-training reviewers were analyzed to develop general recommendations. Additional information specific to Workbook exercises can be found in appendices A, B and C. The following areas of Workbook training are addressed: a) Workbook design; b) Methods of training, and c) Workbook exercises.

Recommendations for Revisions to the Coding Workbook

Workbook Content	Recommendation
Workbook Design	A critical change to the design of the Workbook involves eliminating answers from the protocol to develop an actual "workbook" and answer key. This revision will enable reviewers to complete coding decisions prior to obtaining instructive feedback. General directions for coding should correspond to the revised protocol and instruct reviewers to complete exercises, verify responses, review response explanations and record feedback. Reviewers who used study articles for some exercises found that this method provided information for coding decisions that could not be obtained from study excerpts. It is recommended that the Workbook exercises be conducted with entire research studies. Reviewers that locate study methods for each criterion can verify their decisions on the answer key.
Method of Training	The researcher and secondary reviewer had difficulty with applying knowledge previously learned through statistical courses to the coding tasks. The Coding Workbook, if implemented at the appropriate level of statistics coursework or training workshops, would allow direct application of learned theory. Group discussion could allow for ongoing corrective feedback for reviewers. Through the use of peer-review process, reviewers in this study yielded reciprocal benefits resulting from collaborative problem-solving. As such, a peer-review process that pairs Workbook users at different levels of research experience is recommended as an effective training method.

Workbook Exercises

To assist reviewers with "difficult to code" criteria (as indicated in Appendices A, and B), Workbook exercises should be supplemented with explicit coding instructions and more detailed criteria explanations.

Glossaries that provide detailed definitions of research designs, statistical terms, and other reference topics are recommended. Details for each exercise are listed in Appendices A, B, and C.

Reviewer commentary indicated that additional guidelines for evaluating criteria for Reliability and Validity are needed. Refer to Appendix B for additional information.

Reviewers identified a number of exercises that could not be coded without indications of how to interpret tables, apply formulas, or determine the applicability of statistical methods. Criteria for Statistical Treatment exercise is one such example. For criteria of this type, the Workbook should provide the user with illustrative examples to guide completion of the table or exercise. The Workbook might also list statistical methods that require procedures to control familywise error.

"Correct" responses in the answer key should provide Workbook users with detailed explanations of the process used to arrive at answers.

Results of the case study suggest that additional exercises be added to the Workbook. These additions are recommended to provide reviewers with opportunities to master difficult criteria and to expose users to Manual criteria that are currently not represented in the Workbook.
Specific recommendations for additional Workbook exercises are included in Appendices A and B. In addition, the Workbook should be expanded to include exercises for Type of Program and Stage of Program criteria. Difficulties with applying these criteria to research were noted by reviewers.

As a self-instructional tool, the Workbook has outstanding potential for use in training and practice settings. The results of this case study demonstrated that revisions to the Workbook will be necessary to promote its effective use in graduate training programs. If modified to meet the needs of its consumers, the Workbook could be instrumental in the preparation of school psychologists for evidence-based practice.

Study Limitations

This study was subject to some important limitations. These limitations include the number of studies coded, the inclusion of reviewers at only two levels of training and the length of time between training and study coding.

The Research Study Sample. The sample of research studies reviewed for this project comprised one limitation. Only one study was coded prior to Workbook training. Coding results of article 1 from two of the faculty reviewer and researcher were compared to a post-training review of the same article by the third reviewer. This method limited the comparisons that could be made between pre-and post-training coding outcomes. In addition, the pre-training responses from two reviewers were compared to the post-training responses of another reviewer. Future studies of the Workbook should utilize a larger sample of studies that are similar in research design. Using this method would allow direct comparisons to be made between studies coded by all reviewers prior to and following Workbook training.

The Background of Reviewers. The limited diversity of reviewer's research background was another shortcoming of this study. Only two levels of research experience were represented by reviewers that included a faculty member and two graduate students currently at a Specialist level of training. Therefore, the results of Workbook training for doctoral level graduate students or practicing school psychologists was not examined in this study. The inclusion of graduate students at a post-doctoral level of training would have produced results that were generalized to a larger degree.

The study also failed to include students currently enrolled in statistics courses. These students may have produced different results than the present reviewers due to their concurrent training in statistical analyses. Consequently, there could have been a greater potential for improved coding results and useful reviewer feedback the may have been overlooked due to the similarity of reviewers.

The Time Interval between Workbook Training and Manual Coding. The interval of time between Workbook training and study coding may have influenced the results of this study. Different times for training and study coding were recorded by each reviewer. Therefore, the specific time effects on coding performance are unknown. Reviewers took anywhere from two days to four weeks to complete Workbook training. Subsequent study coding was conducted two weeks after training by one reviewer and 1 month after training by the other two reviewers. Over time, information learned during training may have faded in the memory of reviewers. In addition, coding was conducted in separate locations by each reviewer. The effect these environments may have had on reviewers' coding performance was not examined in this study.

Implications of the Case Study

There are a number of implications associated with this case study. This study highlighted challenges that may be encountered by school psychologists using the Coding Workbook and Procedural and Coding Manual. One implication is suggested by the coding difficulties reflected in post-training results and reviewer feedback. A related implication relates to how the Workbook can be improved to yield positive increases in the coding performance of its consumers. The present study has yielded a number of recommendations intended for this purpose. Revisions to the Workbook that increase its instructive utility will only aid the Task Force in achieving its mission.

Positive results generated in this study also have important implications for use the Workbook. The study demonstrated that repeated study coding sessions are associated with decreases in coding time. This result confirms that assertion by Stoiber (2000) about practice effects and suggests that the Workbook may be useful in training settings. Variability in coding time related to differences in research studies was another outcome of the case study. In addition, positive learning outcomes were generated when reviewers utilized a "peer-review" process that integrated consultative post-coding reviews. Reviewers also stated that Workbook training was conducive to being conducted in manageable learning sessions. Approaches to Workbook coding and results of this case study imply that the Workbook may have utility for providing school psychologists with increased understanding of Manual criteria and empirically-based research.

Conclusion

This case study of the Coding Workbook provides further evidence that the mission adopted by the Task Force is a challenging one. The ultimate goal of Task Force is to promote the application of a scientist-practitioner orientation in practical settings. The Coding Manual and Workbook were developed for this purpose. With the focus of its actual application still to be determined, the Task Force plans to implement the Workbook as a training tool for school psychologists conducting empirically-based research.

The present study examined the practical application of the Workbook in a university setting. Pre- and post-training evidence of reviewers' coding proficiency was analyzed. Detailed observation logs taken by reviewers provided extensive information about their ability to understand Manual criteria throughout all phases of the study. Results of the study suggested that the graduate students lacked the necessary research background to benefit from Workbook instruction and demonstrate an increase in their understanding of the coding criteria. Difficulties with interpreting criteria in Workbook exercises were also recorded by the faculty reviewer. Commentary from reviewers at varied levels of research experience suggests revisions to the Workbook will be necessary in order to increase its usefulness for expanding consumers' knowledge of coding criteria.

Although this study has limitations, it draws attention to the challenges of implementing the Coding Workbook. It also provides recommendations for potential improvements to the Workbook. Determining the effectiveness of the Workbook in promoting school psychologist's ability to evaluate scientific research is important to achieve the overall goal of increasing a scientist-practitioner orientation in schools and other practical settings. To attain this goal, more extensive research is needed to define the unique training needs of school psychologists.

Appendix A

Additional Commentary about Other Descriptive Criteria Exercises

Coding Criteria	Supplementary Information required for Coding
<u>Single Participant Designs</u>:	
External Validity Indicators	An issue was raised with regard to the purpose for having reviewers record a summary of previous ratings.
External Validity Indicators	Reviewers were unsure of the information required for the “Program” section of the response chart.
Length of Intervention	The Workbook protocol was clear and instructive. Although different codes were required for the Manual protocol, the Workbook form was preferred.
Intensity of Intervention	The researcher found the Manual protocol superior to the Workbook protocol format. It required more specific information that was instructive to the process.
Intervention Style	Reviewers were able to code the study excerpt for this exercise accurately. However, theoretical orientation descriptions may be necessary for other school-based studies.
Training and Support Resources	A more specific definition for “simple orientation” and requirements for training workshop criteria would be beneficial to reviewers. The study does not appear to meet criteria for implementation by “typical school staff”. Therefore, responses on the Workbook protocol may be incorrect.

Appendix B

Additional Commentary about General Characteristics Workbook Exercises

Coding Criteria	Supplementary Information for Coding
Single Participant Designs:	
General Design Characteristics	Table 1. Major Types of Single-Participant Design and Associated Characteristics in the Manual provided relevant information about research designs. Reviewers could not distinguish between "simple" and "complex phase change".
Other Design Characteristics (Randomization)	Coders were unclear about use of randomization in the study. The "randomization" in the study appears to refer to the behavioral intervention technique instead of the unit of assignment. Low rates of coding accuracy (average 48 percent) indicate that an additional Workbook exercise is needed for these criteria. Reviewers found these criteria "difficult to code". Reviewers may benefit from a definition or example that clarifies the term "unit of assignment" in B1. The criteria should express "equivalence" in more specific terms.

Group-Based Designs:	
General Design Characteristics	Definitions for Nonrandomized Block Design (between subjects) required further explanation.
Comparison Group	More detailed definitions for Type of Comparison Group found in the Manual were needed to code this criterion. A definition for “intent-to-intervene analysis” is not provided in the Workbook or Manual. Criteria for Equivalent Mortality may require further clarification. Reviewers questioned if attrition should be measured for the entire sample or specifically for each experimental group.
Statistical Treatment	The graduate student reviewers were uncertain of the statistical test used in the study or the method of interpreting Table 1 in the Manual. These reviewers also required additional information to calculate effect size, to determine a sufficiently large N and to evaluate the applicability of controlling for familywise error. Table 1 in the Manual with supplemented explanations of statistic procedures should be added to the Workbook. In addition, reviewers raised questions about methods for determining the Appropriate Unit of Analysis when this information is not available in the research study. These criteria fall under the category of “difficult to code”. Additional Workbook exercises are recommended for Statistical Treatment criteria based on reviewer commentary and study results indicating low rates of coding accuracy. Post-training reviews generated similar results.

Appendix C

Additional Commentary about Evaluation of Key Features Exercises

Coding Criteria	Supplementary Information required for Coding
Single Participant Designs:	
Measurement: Reliability and Validity	The criteria explanation for Validity in version 2 of the Manual is considerably more thorough than the criteria description in the Workbook. Examples of valid measures are also provided. Reviewers stated that information in the Manual was essential for accurate coding decisions. Reviewers stated that further explanation of the term "multi-source" is required. The correct response on the WB protocol should be clarified. The study employed an ABAB Design. Therefore, it is unclear if a coding for randomization is warranted.
Key Outcomes Significant	Chart headings were not easily interpreted by reviewers. Reviewers needed clarification of requirements for "change in level" for treatment phases.
Measures Support Key Outcomes	Graduate student reviewers had difficulty determining responses based on some chart headings (i.e. information required about "Treatment Phases). These criteria were considered "difficult to code" by reviewers.
Educational Significance	The description for D2: Outcome Assessed by Continuous Variables was vague. The description in the Manual was used to clarify this criterion.

Identifiable Components	The Manual contains a coding scheme that is more detailed and informative to the reviewer than that provided by the Workbook.
Replication	The researcher was unclear about whether this criterion referred to replication in the present study or as part of previous research. Also, expectations for researching replication in previous research should be better delineated. One reviewer did not find sufficient evidence for replication in the study excerpt. However, the Workbook protocol indicates the contrary. Clarity of criterion remains necessary. A more detailed definition of study replication would clarify this coding decision for reviewers with less research experience.
Group-Based Designs: Measurement: Reliability and Validity	Examples of "well-known" standardized assessments and more specific criteria were required for coding by the graduate students. The coding for Validity is vague and resulted in multiple interpretations by reviewers. Further clarifications are needed to determine if reviewers should code a "yes" or "no" indication of validity reported in the study or conduct an assessment for validity of study measures.
Key Outcomes	Graduate students lacked the statistical background Significant to interpret Manual Table 1. Additional instructions and examples for calculating effect size were needed. In addition, chart headings (i.e. List Outcome, 1-β) were not easily interpreted by reviewers. These criteria were considered "difficult to code" by reviewers.

	Additional Workbook exercises are recommended for these criteria based on reviewer commentary and study results indicating low rates of coding accuracy. Difficulties with coding Key Outcomes Significant criteria were also recorded during post-training reviews.
Durability of Effects	The reviewers found conflicting results on the Workbook protocol. The boxes indicate a rating of “0” but the protocol comments refer to a rating of “1”.
Implementation Fidelity	The faculty reviewer questioned the need for reporting Acceptable Adherence data as evidence for this criterion. A more detailed explanation of Ongoing Supervision and Case Consultation were required. The criteria might be expanded to inform reviewers about acceptable methods of coding sessions (i.e. the adequacy of an observation method).

Appendix D

Editorial Recommendations for the Coding Workbook Exercises

Coding Criteria	Editorial Suggestions
General Characteristics:	
General Design Characteristics (Single Participant Designs)	Add Table 1 in Version 2 of the Manual entitled *Major Types of Single-Participant Designs and Associated Characteristics* to the Workbook.
Other Design Characteristics	
Randomization (S-P Designs)	The Workbook protocol indicates B3.3 as the correct response but then supplies text supporting a B3.6 coding. The alternative of N/A (randomization not used) for areas B1-B4 seemed extraneous due to other selections that also indicate nonrandom assignment.
Key Features:	
Measurement – Issues of Reliability/Validity (S-P Designs)	Additional examples of studies for which ratings of Multi-method and Multi-source would not apply would be helpful to readers. Differences in the criteria format for Validity between the Manual and Workbook can result in different interpretations. The Manual criteria seem to require an evaluation of the validity of measures that were reported. However, the Workbook requires a coding of whether the validity of measures was reported. Revisions should clarify this point.
Quality of Baseline (S-P Designs)	The study excerpt and graphed data (Theodore et al, 2001) are missing from the exercise.

Key Outcomes Significant (G-B Designs)	The data tables from Fawcett et al. (2001) are missing from the Workbook. The Workbook instructions should make it very clear that a rating of 3 for Measurement and Comparison Group are assumed for practice purposes. If not, reviewers will determine inaccurate codes based on the referenced article excerpt. The answer key provided a useful explanation for the rating of 3 or strong evidence. Further expansion of this information to include the procedure for determining a sufficient large N size would be beneficial to reviewers.
Measures Support Key Outcomes (S-P Designs)	The page for Appendix A is incorrect. The Workbook might instruct reviewers to calculate effect size using Approach 1, provide a chart for doing so and then supply the correct responses. Table 1 (Means, SD and ES) and Figure 1: Percentage of disruptive interval across students) (Theodore et al, 2001) was not illustrated in the Workbook. The exercise references a table on page 22 that was not located by reviewers in the Workbook.
Education Significance (S-P Designs)	The chart for in the Manual *protocol* includes information for outcome variables that do not appear in the Workbook. These descriptions are useful to novice reviewers. For example, Outcomes Assessed via Continuous Variables is defined as a “positive change showing clinical improvement from baseline to intervention). Correct a typographical error in the exercise heading.

Durability of Effects (S-P Designs)	Table 5 from Clarke et al. (2001) study was excluded from the exercise. However, the graphed data in this article that appears to refer to students in reversed order complicated coding. The responses on the Workbook protocol should be revised to reflect the actual number of subjects (4 instead of 5 participants).
Durability of Effects (G-B Designs)	Add table II and III from the Fawcett et al. (2001) study to the exercise. The Workbook protocol indicates a rating of "0" but refers to a rating of '1" in the explanation.
Identifiable Components (S-P Designs)	The exercise lists an inaccurate reference for the study excerpt. Therefore, it must be added to the exercise. Reviewers found information in the Manual protocol coding scheme more useful than the format of the Workbook.
Implementation Fidelity (S-P Designs)	The responses on the Workbook protocol should be revised to reflect the actual number of subjects (4 instead of 5 participants). Thus, the Average Fidelity rating should be .75 instead of .60.
Replication (S-P Designs)	The study by Gettinger (1985) was not included in the exercise.

Other Descriptive Criteria: (Single-Participant Designs)	
Receptivity by the Target Participant Population	Revise the Program and Results topics headings on the Workbook protocol to reflect greater specificity for the information required.
Length of Intervention	Add figure 1 from Theodore et al. (2001) to the exercise.
Characteristics of Intervener	Reviewers suggested the addition of a code for “Unknown/Not Specified”.
Training and Support Resources	The criteria for supports provided *by school or other typical staff* is easily overlooked and misunderstood by reviewers. Thus, this information should be clarified and bolded in the text.

Appendix E

Supplementary Issues Related to Manual Coding

Coding Criteria	Coding Concerns (term definitions, procedural guidelines and explanations of criteria needed)
General Characteristics:	
Statistical Treatment	Guidelines for determining the appropriate level of effect size for coding should be added to address studies that fail to report this factor.
Type of Program (Universal, Selective, Targeted)	The researcher had difficulty determining whether this criterion referred to the intervention model or the implementation in the candidate study.
Stage of Program	Problems with distinguishing between early and established programs were noted by the researcher.
Key Features:	
Measurement	The numerical ratings are unclear. For a rating of "1" the Manual states "In addition data may have been collected either (1) using multiple methods and/or (2) from multiple sources; however, this is not required for a rating of 1". Does this mean that if neither of these cases applies that the study should still receive a rating of "1"? Reviewers recommend adding a code for "information not adequately addressed" to these and other applicable criteria. Coding all "inadequate information" as "0" may result in a non-differentiation between "unknown" and "known, but unsatisfactory".

Comparison Group

A definition for “intent-to-intervene analysis” is not provided in the Workbook or Manual.

Statistically Significant Outcomes

A clarification for the requirement of controlling both family-wise error and experiment-wise error for this criterion is needed. Questions also arose about the acceptability of p-values verified within a Bonferroni adjustment for studies that fail to report methods for controlling family-wise error.

Education Significance

Reviewers were unclear about distinctions between ratings of “No” and “Unknown”. A post-test result for interventions without a categorical diagnosis is one example of this problem. Reviewers debated whether this study should be coded as “no change” or an “unknown” factor.

A more detailed definition for “categorical diagnosis data is needed for distinction between standards of a formal diagnosis and other researcher-defined categories.

Identifiable Components

Reviewers cited the need for greater specificity in the description of “components *linked* to primary outcomes”.

<u>Other Descriptive Criteria:</u>

External Validity Indicators

A rating of N/A might be included to criteria for A1.2 and A1.3 to accommodate studies without inclusion criteria.

Participant Characteristics

The criteria regarding demographics of the sample were interpreted differently by reviewers. The researcher coded information that was indicated for the entire sample. However, the faculty reviewer only recorded demographics that were separately indicated for the intervention and control groups.

Receptivity by the Target Participant Population	Reviewers differed in their perception of the "Target Participant Population". The Manual is unclear about whether these criteria refer to only the treatment subjects or if it also includes the implementers (teachers).
Characteristics of Intervener	A selection for "insufficient information provided" may be added to promote more accurate study coding.
Level of Difficulty in Training Intervention Agents	Measurement standards for "the level of difficulty" should be added to this criterion (i.e. amount of time or perception of difficulty).

Appendix F

Observation Log Form

Name:

Instrument: □ Coding Workbook □ Coding Manual

Session 1: 1/00/05	Session 2: 1/00/05	Session 3: 0/00/05	Session 4: 0/00/05
Start time: End time:	Start time: End time:	Start time: End time:	Start time: End time:
Session 5: 1/00/05	Session 6: 1/00/05	Session 7: 0/00/05	Session 8:
Start time: End time:	Start time: End time:	Start time: End time:	Start time: End time:

Enter comments regarding each section of the Coding Workbook below:

Part I: General Characteristics

General Design Characteristics (Single Participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

General Design Characteristics (Group-based Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

General Design Characteristics and Comparison Group (Group-based Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Other Design Characteristics – Randomization is Used (Single Participant Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Statistical Treatment (Group-based Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Part II: Key Features

Measurement-Issues of Reliability and Validity (Single Participant Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Measurement (Group-based Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Quality of Baseline (Single-participant Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Key Outcomes Significant (Group-based Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Measures Support Key Outcomes (Single-participant design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Educational/Clinical Significance (Single-participant design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Durability of Effects (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Durability of Effects (Group-based Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Identifiable Components (Single-participant Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Implementation Fidelity (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Implementation Fidelity (Group-participant Designs):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Replication (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Site of Implementation (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Part III: Other Descriptive or Supplement Criteria to Consider

External Validity Indicators – Summary of Key External Validity Indicators (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

External Validity Indicators – Participant Selection and Description (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

External Validity Indicators – Receptivity/Acceptance by Target Participant Population (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Length of Intervention (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Intensity/Dosage of Intervention (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Program Implementer (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Intervention Style or Orientation (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Cost-Benefit Data (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

Training and Support Resources (Single-participant Design):
(Record general comments about the Workbook criteria and coding exercise).

Assess the effectiveness of information in the Workbook response key:

Please make suggestions for improvements to the above section (article excerpt, Manual criteria, coding protocol or the illustrative example in general):

References

Ahn, H., & Wampold, B. E. (2001). Where oh where are the specific ingredients: A meta-analysis of component studies in counseling and psychotherapy. *Journal of Counseling Psychology, 48(3),* 251-257.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association, Division 12 (1993). Task Force on promotion and dissemination of psychological procedures. Retrieved May 6, 2004 from http://www.apa.org/divisions/div12/journals.html.

Behring, S. T., & Ingraham, C., L. (1998). Culture as a central component of consultation: A call to the field. *Journal of Education and Psychological Consultation,* 9(1), 57-72.

Beulter, L.E. (1998). Identifying empirically supported treatments: What if we didn't? *Journal of Consulting and Clinical Psychology, 66(1)* 113-120. (special issue).

Carnine, D. (2002). National Center to Improve the Tools of Educators. Eugene, OR: University of Oregon.

Center for Disease Control (1999). Guidelines for National Human Immunodeficiency Virus Case Surveillance, Including Monitoring for Human Immunodeficiency Virus Infection and Acquired Immunodeficiency Syndrome. Retrieved June 20, 2004 from (http://www.cdc.gov/nccdphp/dash/rtc/index.htm).

Chambless, D. L. (2002). Identification of empirically supported counseling psychology interventions: Commentary. *Counseling Psychologist. 30* (2), 301-308.

Chambless, D. L., Baker, M. J., Baucom. D. H., Beutler, L E., & Calhoun, K. S. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist, 51(1),* 3-16.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66(1),* 1-18.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology. 52*, 685-716

Christenson, S. L., Carlson, C., & Valdez, C. R. (2002). Evidenced-based interventions in school psychology: Opportunities, challenges, and cautions. *School Psychology Quarterly, 17,* 466-474.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Counseling and Clinical Psychology,* 59, 20-26.

Durlak, J. A. (2002). Evaluating evidence based interventions in school psychology. *School Psychology Quarterly, 17(4)*, 475-482. (special issue).

Fals-Stewart , W., Marks A.P., & Schafer J. (1993). A comparison of behavioral group therapy and individual behavior therapy in treating obsessive-compulsive disorder. *Journal of Nervous Mental Disorders* 181:189–93

Garfield, S.L. (1998). Some comments on empirically supported treatments. *Journal of Consulting and Clinical Psychology, 66(1),* 121-125. (special issue).

Greenberg, M.T., Kusche, C.A,, Cook, E. T., & Quamma, J. P. (1995). Promoting emotional competence in school-aged children: The effects of the PATHS curriculum. *Development and Psychopathology, 7,* 117-136.

Gutkin, T. B. (2002). Could it be over soon? *School Psychology Quarterly,* 17(4), iii-v. (special issue).

Gutkin, T. B. & Curtis, M. J. (1999). School-based consultation theory and practice: The art and science of indirect service delivery. In C. R. Reynolds & T. B. Gutkin (Eds.), *The Handbook of School Psychology,* (598-637), New York: John Wiley.

Haney, P., & Durlak, J. (1998). Changing self-esteem in children and adolescents: A meta-analytic review. *Journal of Clinical Child Psychology,* 27(4), 423-433.

Henry, W. P. (1998). Science, politics and the politics of science: the use and misuse of empirically validated treatment research. *Psychotherapy Research, 8(2)*, 126-140.

Hibbs, E. D. (2001). Evaluating empirically based psychotherapy research for children and adolescents. *European Child & Adolescent Psychiatry. 10*, 1-11.

Hughes, J. N. (2000). The essential role of theory in the science of treating children. Beyond empirically supported treatments. *Journal of School Psychology, 38(4),* 301-330.

Kadzin, A. E., & Weisz, J. R. (1998). Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology, 66,* 19-36.

Kam, C., Greenberg, M. T., & Walls, C.T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4(1),* 55-63.

Kendall, P. C. (1998). Empirically supported psychological therapies. *Journal of Counseling Clinical Psychology,* 66(1), 3-6. (special issue).

Kratochwill, T. R. (2002). Evidence-based interventions in school psychology: Thoughts on a thoughtful commentary. sc*hool Psychology Quarterly, 17(4),* 518-532 (special issue).

Kratochwill, T.R. & Shernoff, E. S. (2004). Evidence-Based Practice: Promoting evidence-based interventions in school psychology. *School Psychology Review,* 33, 34-49.

Kratochwill, T.R., & Stiober K. (2002) Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force, *School Psychology Quarterly, 17(4),* 341-389 (special issue).

Kroesbergen, E. H., & Van Luit, J. (2003). Mathematics interventions for children with special education needs: A meta-analysis. *Remedial & Special Education, 24(2*), 97-114.

Kusche, C. A. & Greenberg, M. T. (1994). *The PATHS Curriculum.* Seattle, WA: Developmental Research and Programs.

Larson, K. A. (1989). Task-related and interpersonal problem-solving training for increasing school success in high-risk young adolescents. *Rase, 10(5),* 32-41.

Lehr, C. A., Hansen, A., Sinclair, M. F., & Christenson, S. L. (2003). Moving beyond dropout towards school completion: An integrative review of data-based interventions. *School Psychology Review, 32*, 342-364.

Lentz, F. E., Allen, S. J., & Ehrhardt, K, E. (1996). The conceptual elements of strong interventions in school settings. *School Psychology Quarterly, 11(2),* 118-136.

Levant, R.F. (2004). The empirically validated treatments movement: A practitioner/educator perspective. *Clinical Psychology: Science and Practice, 11(2)* 219-224.

Levin, J. R. (2002). How to evaluate the evidence of evidence based interventions? *School Psychology Quarterly*, *17(4),* 483-492. (special issue).

Lewis-Snyder, G., Stiober, K. C., & Kratochwill, T. R. (2002). Evidence-based intervention in school psychology: An illustration of the task force coding criteria using group-based design. *School Psychology Quarterly, 17(4),* 423-465. (special issue).

Lewis, T. J, & Sugai, G. (1999). Effective Behavior Support: A systems approach to proactive school-wide management. *Focus on Exceptional Children, 31(6),* 1-24.

Lindsay M, Crino R, Andrews G. 1997. Controlled trial of exposure and response prevention in obsessive-compulsive disorder. *Br. J.Psychiatry* 171:135–39

Loeber, R., Wei, E., Stouthamer-Loeber, M. (1999). Behavioral antecendents to serious and violent offending: Joint analysis from the Denver Youth Survey, Pittsburgh Youth Study and the Rochester Youth Development Study. *Studies on Crime and Prevention. 8(2)*, 245-263.

Mayer, M. J., & Leone, P. E. (1999). A structural analysis for school violence and disruption: Implications for creating safer schools. *Education & Treatment of Children, 22(3),* 333-347.

McGuire, J. (1985). Methodological quality as a component of meta-analysis. *Educational Psychologist. 20(1),* 1-5.

National Association of School Psychologists. (2001). Zero tolerance and alternative strategies: A fact sheet for educators and policymakers. Retrieved July 3, 2004 from www.naspcenter.org.

National Education Goals Panel. (2000*). Promises to keep: Creating high standards for American students.* Washington, DC: Author.

Nelson, J. R. & Epstein, M. H. (2002). Report on evidence based interventions: Recommended next steps. *School Psychology Quarterly, 17(4)*,493-499. (special issue).

Norcross, J. C. (2001). Purposes, processes, and products of the Task Force on Empirical Therapy Relationship. Psychotherapy: Theory/Research/ Practice/Training*, 38,* 345-356.

OSEP Center on Positive Behavior Interventions and Supports. (2001). School-wide Positive Behavior Support. Retrieved on July 22, 2004 from http://www.pbis.org/english/Schoolwide PBS.htm.

Prevatt, F., and Kelly, F. D. (2004). Meeting the challenge of identifying evidence based interventions in school psychology. Manuscript submitted for publication.

Ramirez, S. J., Lepage, K. M., Kratochwill T. R., & Duffy J. L. (1998). Multicultural issues in school-based consultation: Conceptual and research considerations. *Journal of School Psychology*. *36(4),* 479-509.

Reed, G. M., McLaughlin C., & Newman, R. (2002). American Psychological Association policy in context: The development and evaluation of guidelines for professional practice. *American Psychologist, 57,* 1041-1047.

Reschlt, D. J. (2000). The present and future status of school psychology in the United States. *School Psychology Review, 29,* 507-522.

Safran, S. P., & Oswald, K. (2003). Positive behavior supports: Can schools reshape disciplinary practices? Exceptional Children, 69(3). 361-374.

Sarason, S. (1991). Revisiting *"The culture of the school and the problem of change."* N.Y.: Teachers College Press.

Seifer, R., Gouley, K., Miller, A. L., & Zakriski, A. (2004). Implementation of the PATHS curriculum in an urban elementary school. *Early Education and Development, 15(4),* 485.

Sheridan, S. M., Welch, M., & Orme, S. F. (1996). Is consultation effective? *Remedial and Special Education,* 17(6), 341-354.

Shernoff, E. S., & Kratochwill, T.R. (2003). *Coding Single-Participant and Group Design Studies: Coding Workbook for Evidence-Based Interventions*. Unpublished Manual.

Shernoff, E. S., Kratochwill, T. R. & Stoiber, K. C. (2003). Training in evidence-based interventions (EBIs): What are school psychology programs teaching? *Journal of School Psychology, Vol. 41(6),* pp. 467-483.

Shernoff, E. S., Kratochwill, T.R., & Stoiber, K.C. (2002). Evidence-Based interventions in school psychology: An illustration of task force criteria using single-participant design. *School Psychology Quarterly, 17(4),* 390-422 (special issue).

Shinn, M. R., Walker, H. M, & Stoner, G. (EDS.). (1991). *Interventions for academic & behavior problems II: Preventive and remedial approaches.* Bethesda, MD: National Association of School Psychologists.

Stoiber, K. C. (2002). Revisiting efforts on constructing a knowledge base of evidence-based intervention within school psychology. *School Psychology Quarterly,* 17,533-546. (special issue).

Stoiber, K. C. & Kratochwill, T. R. (2000). Empirically supported interventions and school psychology: Rationale and methodology issues – *Part I. School Psychology Quarterly, 15,* 75-105.

Stoiber, K. C. & Waas, G. A. (2002). A contextual and methodological perspective on the evidence-based intervention movement with school psychology in the United States. *Educational and Child Psychology*, *19(3).*

Sue, S. (1999). Science, ethnicity, and bias: Where have we gone wrong? *American Psychologist, 54,* 1070-1077.

The Evidence-Based Intervention Work Group (2005). Theories of change and adoption of innovations: The evolving evidence-based intervention and practice movement in school psychology. *Psychology in the Schools, 42,* 475-494.

The Task Force for Evidence-Based Interventions in School Psychology (2003). *Procedural and Coding Manual for the Review of Evidence-Based Interventions, version 2*. Washington, DC: The American Psychological Association.

Wampold, B. E. (2002). An examination of the bases of evidence based interventions. *School Psychology Quarterly, 17,* 500-507. (special issue).

Wass, G. A. (2002). Identifying evidence based interventions in school psychology: Building a bridge or jousting with windmills? *School Psychology Quarterly, 17,* 508-517. (special issue).

Weisz, J.R. & Hawley, K.M. (1998). Finding, evaluating, refining and applying empirically supported treatments for children and adolescents. *Journal of Clinical Child Psychology, 27*, 206-216.

What Works Clearing House (2004). Overview of What We Do. Retrieved October 30th from www.w-w-c.org/whatwedo/overview.html.