THE EFFECTS OF TRAINING IN PHONEMIC AWARENESS: A META-ANALYSIS

by

Jan Brewer Miller
B.A., Birmingham-Southern College, 1989
M.S., University of Alabama, 1993

Submitted to the Department of Teaching and Leadership and the Faculty of the Graduate School of the University of Kansas in partial fulfillment of the requirements for the degree of Doctorate of Education.

ABSTRACT

This study quantitatively examined the relationship between phonemic awareness training and the phonemic awareness of the trained subjects. Eighteen studies were coded for continuous and categorical moderator variables.

The overall weighted mean effect was strong. Differences between the subgroups within Approach of Training were significant, favoring segmenting/blending training. Differences in Outcome Measures were also significant. The strongest effect size for Outcome Measures was Combination Measures.

Minutes per Training Session and Total Minutes in Training were negatively and significantly correlated with effect sizes. As training time increased, effect sizes tended to decrease. Group Size was inversely and significantly correlated with phonemic awareness. After controlling for group size, Total Minutes in Training remained negatively and significantly related to effect size.

For Stuart

ACKNOWLEDGEMENTS

My twin sons and this paper were born around the same time; the labor for the latter has been longer and more exhausting. The boys are beginning preschool as I am finishing this project; my children and my dissertation have grown-up together, and I with them. Now I sit attempting to capture on paper my gratitude to the people who nurtured me. However, I am far more grateful than I am eloquent.

I jumped into this project without looking to see how far I had to fall or exactly where I would land. Fortunately, I found myself caught by a network of colleagues and family who were there for the initial plunge and for the years of stumbling that followed. It is relatively easy to learn to walk when there is always a hand to hold. For their faithful execution of hand-holding duties I am eternally grateful to the following:

- Dr. Nita Sundbye, my first advisor, who originally discussed with me the excess of phonemic awareness research and possibilities for summarizing it. Fortunately for me, she was the first person I met at the University of Kansas and I was one of her last doctoral students.
- Dr. Diane Nielsen, my second advisor, who never allowed me to get by with too little and never pushed me to a point of frustration. Diane's frankness, character, kindness, and persistence have been models for me both personally and professionally.

- Dr. Sam Green, my statistics professor, who donated countless hours to laboring through my early drafts and to answering my endless questions. He has been caring, attentive, exacting, and patient.
- Dr. Steve White and Dr. Arlene Barry who have been flexible, supportive, and kind throughout my doctoral experience.
- Dr. Marilyn Thompson, who was the leader in our doctoral game of Follow-the Leader. It was always helpful that she did things first and well. My appreciation for Marilyn runs deeper than the pages of this paper and my pen cannot capture it.
- Dr. Blair Johnson, a meta-analysis expert, who entered this dissertation story at the climax, when my tension was high and I doubted I would ever see it completed. He was the resolution of this drama and he was kind enough to chew far more than he bit off.
- My parents, all three of them, who first taught me to nurture my curiosity and to set my heart and mind on what is important.

Finally, I am most indebted and most grateful to my husband, Stuart, who has been unflinchingly selfless throughout this project. He shouldered every responsibility that I released as I worked on my dissertation and nudged me forward with his words of encouragement. All of my outpourings of gratitude are grossly inadequate in expressing my appreciation of Stuart. He has been my great enabler.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Statement of the Problem

The early 1900s marked the beginnings of reading research. These studies were physiological and psychological in nature, focusing on eye-movement and visual perception. By 1910 there were only 34 studies in the United States and England that dealt with reading. The early 1900s brought the advent of instruments of measurement which made it possible to gather scientific information about the effectiveness of reading methods and materials. In the years between 1910 and 1925, reading research expanded rapidly under the influence of standardized testing. By 1925, there were 426 research studies on topics such as silent reading, reading rate, phonics, and grouping for reading (Smith, 1974).

Now, 71 years later, the same topics are investigated. The cumulative effect is volumes of research on many reading-related subjects. To understand the expanse of information, researchers assemble reviews of the literature. As scholars cannot know in detail the data in more than a few areas, they often rely on these qualitative summaries of research, even though such reviews are plagued with shortcomings.

Critics of traditional methods of summarizing research have been vocal (see Jackson, 1980). In the tide of these criticisms, Glass (1976) pioneered the meta-analysis, a method for quantitatively summarizing studies. Many researchers have touted meta-analysis as a more objective way to discover the cumulative knowledge in a particular area.

This chapter defines the terms central to the present analysis. Through a brief summary of a portion of the literature on phonemic awareness, this chapter also attempts to illustrate the broad scope of the phonemic awareness literature. The summary is followed by a discussion of the limitations of narrative reviews. Finally,

the need for this study is established based on the preponderance of phonemic awareness data and the shortcomings of traditional narrative summaries. The chapter closes with a list of research questions.

Definitions

Before analyzing data, the definitions of phonemic awareness and phonological awareness must be clarified. Phonemic awareness, which is currently the term reading researchers most commonly use in studies, is very specific; however, there is some confusion over the terms phonemic awareness and phonological awareness. Phonological awareness includes word awareness, syllable awareness, rhyme awareness, and phonemic awareness. Phonemic awareness deals with the awareness and manipulation of the smallest units of speech, phonemes. Phonemic awareness tasks include phoneme segmentation, blending, deletion, addition, and substitution.

The body of phonemic awareness research provides some well-articulated definitions of phonemic awareness. Phonemic awareness, as defined by Stanovich (1986) is "conscious access to the phonemic level of the speech stream and some ability to cognitively manipulate representations at this level" (p. 362). Cunningham (1990) simply stated, "Phonemic awareness is the ability to explicitly manipulate speech segments at the phoneme level" (p. 429). While definitions of phonemic awareness varied in wordiness and tone, most maintained two main ideas: phonological awareness deals with phonetics rather than graphemics and phonemic awareness involves the explicit manipulation of those phonemes.

Research on Phonemic Awareness

The following narrative presentation taps into the diversity of the research on phonemic awareness by investigating four areas: interpretations of the role of

phonemic awareness, recommendations for instruction in phonemic awareness, implications of phonemic training for disabled readers, and insights into assessing phonemic awareness. In order to further clarify the differences in phonemic awareness and phonological awareness, the occurrence of each in the following section is noted in parentheses. Studies included in the statistical analysis are indicated with an asterisk (*).

The Role of Phonemic Awareness in Learning to Read

The debate over whether phonemic awareness is a result or a cause of learning to read is at the forefront of investigations into the role of phonemic awareness. On this issue, Stahl and Murray (1994) concluded that after students achieve an adequate level of letter recognition, the ability to manipulate onsets and rimes within syllables (phonological awareness) is most strongly related to reading achievement. However, Stahl and Murray also found that the ability to isolate a phoneme from either the end or the beginning of a word (phonemic awareness) is also critical to beginning reading because almost all of the subjects who could not perform this task adequately could not read preprimer materials.

Similarly, *Fox and Routh (1984) trained children in analysis and blending (phonemic awareness) and then gave them a reading analog test. This involves training subjects to identify letter-like forms in association with authentic phonemes. These forms are strung together to form real words from novel graphemes. The findings of Fox and Routh suggested that phonemic awareness skills are causally related to learning to read and are not just a by-product of acquired reading ability. *Cunningham's (1990) experimental study confirmed the findings of Fox and Routh (1984). She stated,

> The results are inconsistent, however, with the hypothesis that phonemic awareness is just a consequence of learning to read. If phonemic awareness was simply a by-product of reading ability, then training studies or prior knowledge would have no effect on the development of reading achievement. ...The results of the present study do not, however preclude the possibility that reading influences phonemic awareness. (p. 440)

Bryant, MacLean, Bradley, and Crossland (1990) investigated the relationship between phoneme detection and rhyme awareness and their bearing on a child's beginning reading success. They found that rhyme awareness (phonological awareness) is both directly and indirectly related to reading. Directly, "rhyme and alliteration definitely make an independent and distinctive contribution to reading" (p. 432). Indirectly, recognition of rhyme and alliteration act as a building block for future skill in phoneme detection.

Most of the studies cited here included statements affirming the idea that phonemic awareness has predictive value for beginning reading success. There is a chicken-and-egg question related to the acquisition of reading skills and phonemic awareness. While the research indicates that phonemic awareness influences later reading ability, it is unclear to what degree beginning reading abilities encourage phonemic awareness. Perhaps Perfetti (1984) made the most sense of this puzzle when he said that "a child's learning to read is helped by the emergence of some levels of phonemic awareness and that deeper levels of phonemic awareness may be a consequence of learning to read" (p. 51).

Instruction in Phonemic Awareness

Having established definitions of phonemic awareness and the role of phonemic awareness in beginning reading, the following is a discussion of

instructional recommendations that have grown from research. *Byrne and Fielding-Barnsley (1991) found that preschool children can be successfully trained to identify phonemes. They also found that once children acquire the construct of phoneme identity, they can transfer this knowledge to unknown phonemes. Thus direct instruction of each phoneme is unnecessary if the educational objective is detecting like phonemes between words.

*Fox and Routh (1984) found that children who were trained in segmenting and blending were more successful at a reading analog task than were children who were only taught segmenting. These results suggest that phonemic awareness training should encompass both of these opposite processes. This finding was confirmed by *Torgesen, Morgan, and Davis (1992) who found that blending skill taught without segmenting skill was less effective than when the two skills were taught together. The researchers stated, "The difference in speed of acquisition of new word pronunciations suggests that children in the AB [analysis and blending] group were able to generalize the oral-language phonological awareness they had acquired in training to a novel task: learning to read new words" (p. 369). Thus the evidence of both studies suggested that training in phonemic awareness should encompass analysis and blending.

While most discussions of phonemic awareness and reading ability deal with word recognition, Spedding and Chan (1993) investigated the relationship of phonemic awareness and comprehending abilities. They found that the phonemic awareness ability of phoneme deletion and the strategic use of phonic clues influence reading comprehension indirectly through their effects on blending skills. They suggest that efforts toward reading remediation which have focused on blending training have offered contradictory results. The results of Spedding and Chan's study

implied that problems in blending may actually be a reflection of deficiencies in the underlying skill of distinguishing sounds in words. Spedding and Chan further explained that skill in blending must precede automaticity in word recognition, a prerequisite for comprehension.

Griffith, Klesius and Kromrey (1992) compared the decoding skills, spelling skills, and writing fluency of children with varying levels of phonemic awareness in whole language and traditional instruction classrooms. They found that children who entered the class with a higher level of phonemic awareness consistently outperformed the children who entered with a low level of phonemic awareness.

> That the children in the WLI [whole language instruction] classroom could read nonsense words at a level equal to that of children in the TI [traditional instruction] classroom is an important finding because this test [test of nonsense words] was particularly insensitive to the strategy they had been taught. (p. 90)

Thus, Griffith and her colleagues recommended that children who are low in phonemic awareness be explicitly trained in hearing the sounds in words before they enter first grade.

*Cunningham (1990) also explored explicit versus implicit training in phonemic awareness. Both groups were given direct instruction in phoneme identity. The explicit group also received instruction that emphasized the application, utility, and value of phonemic awareness in learning to read. Cunningham found the latter method to be significantly more effective than the former. She also found that trained kindergartners outperformed untrained first graders. This suggests that after a certain age, development may be less critical than training for children to develop phonemic awareness skills. Cunningham's results are compatible with those of Griffith et al.;

they both found positive results from training kindergartners in phonemic awareness before they enter first grade.

Phonemic Awareness Training for Disabled Readers

Logic has led researchers to pursue the idea that if there is a causal relationship between phonemic awareness and learning to read, then perhaps students with reading difficulties have deficits in phonemic awareness. Hurfod and Sanders (1990) studied the phonemic awareness of reading disabled second and fourth graders. They found that disabled readers in second grade were significantly less proficient than their nondisabled counterparts at discriminating between combinations of two-syllable phoneme pairs, such as /bi/, /di/, and /gi/, with varying intervals between pairs. However, fourth grade disabled readers were as skilled at this task as their nondisabled grademates. Hurford and Sanders concluded that these results indicated that the second graders had a developmental lag in phonemic awareness.

In a second experiment, Hurford and Sanders (1990) investigated the possibility that the children who performed poorly in their first experiment could improve their performance with training. The researchers trained the subjects in tasks similar to those in the first experiment. The training lasted for 30 to 45 minutes for three days. Hurford and Sanders found the performance of the subjects significantly improved after the training.

Williams, (1980) studied classrooms of learning disabled children, ages seven to twelve, who were participating in an instructional program called "The ABD's of Reading". The program provided explicit training in phoneme analysis, phoneme blending, letter-sound correspondences, and decoding. Upon completion of training, the experimental group performed significantly better than the control group on measures of first letter-sound correspondence, second letter-sound correspondence,

saying the middle phoneme, saying all three phonemes, and decoding. In addition, the experimental group performed significantly better on transfer tasks.

O'Conner, Jenkins, Leicester, and Slocum (1992) examined the feasibility of teaching phonemic awareness skills to children enrolled in a special education preschool. Children trained in rhyming, blending or segmenting performed significantly better than students in the control condition. However, the subjects were unable to generalize between tasks within a particular category of training, such as generalizing from one type of blending task to another type of blending task. They were also unable to generalize between categories of tasks, such as generalizing from blending to segmenting.

Minus (1992) examined the phonological awareness of adult disabled readers. She identified 19 inmates from a correctional facility who demonstrated deficits in phonemic awareness. These subjects were trained in syllable segmentation, alliteration, rhyme, blending, phoneme counting, and deletion. Subjects experienced little or no increase in decoding, word recognition, or phonological awareness after twelve weeks of instruction. The author concluded that phonological awareness may be difficult to teach adults.

Assessing Phonemic Awareness

Finally, Spector (1992) suggested that we not only examine how we teach children phonemic awareness but also examine our methods of assessing it. She hypothesized and found that a dynamic assessment would provide a more accurate prediction of beginning reading success. A dynamic measure emphasizes both the process and the product of the assessment, giving students feedback on their performance during the evaluation. Spector proposed that "poor performance on a phoneme segmentation task might indicate low phonemic awareness, it might also

reflect the child's lack of understanding of task requirements or difficulty in meeting ancillary task demands" (p. 1). In addition, Spector found that the students who made the most progress in word recognition during the kindergarten year were the students who benefited from the prompts and instructional cues during the dynamic testing session.

Bryant, MacLean, Bradley, and Crossland (1990) also studied assessment. They found that the relationship between rhyme and reading is specific for phoneme deletion tasks. However, there is a high correlation between phoneme tapping tasks and students' mathematical skill. The authors suggested that this may indicate that the phoneme tapping task is not a pure test of phonemic awareness because it involves counting. Nevertheless, Bryant, MacLean, Bradley, and Crossland confirmed, as have many others, the existence of "a strong, consistent, and specific relation between children's phonological skills and reading" (p. 437).

Conclusion

The purpose of the previous presentation of the research on phonemic awareness training was to illustrate the variety within the literature, rather than to detail a comprehensive narrative summary. The scope of the discussion included the relationship of phonemic awareness to beginning reading, instruction in phonemic awareness, phonemic awareness training for disabled readers, and phonemic awareness assessments.

Need for the study

Preponderance of Data

The studies described in the previous paragraphs are only an initiation to the body of research proporting a relationship between phonemic awareness and beginning reading success. The published studies supporting this relationship are

practically limitless, spanning over thirty years (e.g., Bradley & Bryant, 1983; Calfee, Lindamood, and Lindamood, 1973; Chall, Roswell, & Blumenthall, 1963; Elkonin, 1973; Fox & Routh, 1976; Helfgott, 1976; Juel, 1988; Juel, Griffith, & Gough, 1986; Liberman, Shankweiler, Liberman, Fowler, & Fischer, 1977; Lomax & McGee, 1987; Lundberg, Olofsson, & Wall, 1980; Stanovich, Cunningham, & Feeman, 1984; Swank & Catts, 1991; Tunmer, Herriman, & Nesdale, 1988; Tunmer & Nesdale, 1985; Zifcak, 1981). The body of data is so redundant and compelling that the relationship between phonemic awareness and reading is widely accepted by the reading community.

Almost all researchers agree on the critical relationship between learning to read and skill in phonemic awareness. Researchers also tend to agree that children can be successfully trained in phonemic awareness. However, researchers have not established which training and assessment variables lend the greatest effects in phonemic awareness.

Phonemic awareness has unique appeal to reading researchers. It is the nexus where holistic theorists and skills-based theorists meet. Thus, wherever researchers fall on the "whole-language/phonics" continuum, however flawed these labels may be, they are likely to have high interest in phonemic awareness. Phonemic awareness is the one thing that purists on either extreme agree is important, and they all have been producing studies (Griffith, et al, 1992; Williams, 1980; Winsor, 1990).

In addition, for every academician who has statistically investigated phonemic awareness and concluded by saying, "I found A, B, and C, but more research is needed in this area," researchers have heeded the call and added study upon study to the morass of data relating to phonemic awareness. Rosenthal (1978) suggested that

we are better at responding to the call for more research than we are at figuring out what to do with the answers. Such has been the case with phonemic awareness.

Limitations of Narrative Reviews

Narrative reviews of the literature on phonemic awareness exist, both in their own right and in summaries of reading research, (e.g., Adams, 1994; Bryen & Gerber, 1987; Elbro, 1996; Nicholson, 1996; Smith, 1995; Wong, 1986). However, unless readers are willing to locate and study all of the original research, they will not be able to verify the objectivity of the reviewer. It is easy for even the most skilled narrators and summarizers to unwittingly misrepresent information (Chall, 1983).

Historically, narrative reviews of research literature have attempted to help scholars and policy makers draw conclusions from mounds of data, but subjectivity within reviews makes them irreplicable. Cook and Leviton (1980) defended the narrative review, maintaining that criticisms of qualitative reviews are based more on poor techniques of the reviewers than on the review method itself. In fact, Cook and Leviton argued that

> While qualitative reviews may be equally prone to bias, the descriptive accuracy of a point estimate in a meta-analysis can have mischievous consequences because of its apparent "objectivity", "precision", and "scientism." To naive readers, these lend a social credibility that may be built on procedural invalidity. (p. 455)

Indicting narrative reviews and defending quantitative reviews, Glass, McGaw, and Smith (1981) explained that absorbing the essence of a hundred research studies is as unlikely as being able to scan a hundred test scores without the use of statistical methods. In addition, Cooper (1982) stated that researchers cannot take the conclusions drawn by narrative reviewers at face value. Chall (1983) further stated, "I have been struck by how easy it is to misinterpret findings. The best of us can be led into making hasty conclusions and overgeneralizing from limited evidence" (p.

87). Research consumers must recognize that the integration of individual studies involves "scientific inferences as central to the validity of knowledge as the inferences made in primary data interpretation" (Cooper, 1982, p. 291).

Jackson (1980) investigated the quality of narrative techniques for reviewing research. He was unable to locate a strong body of literature describing the specifics of the methodology. He examined 36 research reviews from well-respected periodicals in the social sciences. His investigation exposed problems with methods for articulating the hypothesis; sampling and including studies; and describing, analyzing, interpreting, and reporting the results from original studies. Jackson concluded that the quantitative review should be structured like a primary study and should replace the narrative review.

An additional Achilles heel for narrative reviews is that the data on the topic under review may be so voluminous that the qualitative reviewer may be forced to exclude studies. It is not unlikely that decisions for exclusion will be made along the lines of the researcher's bias. Narrative reviewers often pay considerable attention to studies that fit the current educational views and only slight attention to studies that contradict these views. Chall (1983) addressed these problems, which can introduce bias to reviews, in Learning to Read: The Great Debate. She stated,

> One of the most important things, if not *the* most important thing I learned from studying the existing research on beginning reading is that it says nothing consistently. . . . And if you select judiciously and avoid interpretations, you can make the research "prove" almost anything you want it to. (p. 87)

Studies may also be excluded from meta-analyses. However, in a quantitative summary, exclusion criteria are more clearly defined. Cook and Leviton (1980)

concede that when there is a large number of studies to be synthesized, meta-analysis has a clear advantage over qualitative reviews.

The previous discussion attempted to demonstrate that the studies exploring phonemic awareness are voluminous and that the traditional methods of summarizing literature are limited in their capacity to authentically represent a large body of literature. Given these two points, a quantitative summary of the phonemic awareness training literature and its effects on phonemic awareness was conducted.

Purpose of the Study

The purpose of this study was to apply meta-analytic techniques to a particular body of the research on phonemic awareness training. This meta-analysis addressed three research questions: 1) Does phonemic awareness training affect phonemic awareness? 2) Which, if any, categorical training variables affect phonemic awareness? 3) Which, if any, continuous training variables affect phonemic awareness?

Conclusion

This chapter argued the need for this study. The presentation articulated definitions that guided this investigation and elaborated on them in the context of a brief literature summary. Next, it maintained that the abundance of phonemic awareness training research could best be summarized quantitatively. Based on this conclusion, a meta-analysis examining the relationship between phonemic awareness training and phonemic awareness was conducted.

CHAPTER II

Review of the Literature

Because the objective of a meta-analysis is to review the literature, the traditional literature review section of this study was treated differently. This chapter serves three purposes. First, it makes a case for the importance of literature reviews, whether quantitative or qualitative. Next, it describes the steps in conducting a meta-analysis, beginning with defining the question and ending with the interpretation of results. Finally, this chapter presents several examples of quantitative reviews conducted in the field of reading.

The Need for Research Reviews

Scientific investigations are cumulative by nature. However, today more than ever, the job of considering accumulated research is a troublesome endeavor. Understanding the investigations that have preceded us is referred to by Pillemer (1980) as the science of discovering what we already know. We are scientifically outrunning ourselves in our research endeavors, making it virtually impossible to comprehend all that we have learned. For example, in 1917 there were only three doctoral dissertations on reading (Smith, 1974). In 1996, there were 19, 811 dissertations dealing with reading and education referenced in Comprehensive Dissertation Abstracts. Furthermore, there are 750 journals containing more than 500,000 articles referenced in the Educational Resources Information Center (ERIC) (Houston, 1995).

It is becoming increasingly difficult to keep up with what we are learning. What may seem clear with one study can become considerably less focused with twenty. Ten studies that found a relationship significant will be thrown into question by one study which found the same relationship nonsignificant. The future of

research depends heavily on our ability to synthesize the ever-expanding literature on a topic, although this has become quite complicated.

Light and Smith (1971) discuss the contradictory nature of scientific inquiry. They explain that inconsistencies in research can be disconcerting when purposes are theoretical in nature. However, when reviews are a source of guidance for those developing public policy, inconsistencies can be paralyzing. In an address to the American Educational Research Association, former Senator Walter Mondale articulated this point in reference to school integration:

> "I have found very little conclusive evidence. For every study, statistical or theoretical, that contains a proposed solution or recommendation, there is always another, equally well documented, challenging the assumptions or conclusions of the first. No one seems to agree with anyone else's approach. But more distressing: no one seems to know what works. As a result I must confess, I stand with my colleagues confused and often disheartened." (cited in Light & Smith, 1971, p. 431)

Blimling (1988) suggested two problems with the absence of a clear, cumulative portrait of research in a field. First, there is no agreement on what research in a particular area reveals. This leaves scholars, policy makers, and practitioners, endlessly debating the contributions of individual studies. A second problem is that researchers put energy into repeating existing knowledge rather than extending it.

Meta-analysis

History of Quantitative Reviewing

The earliest attempts at uniting research studies in a systematic way involved locating theoretically relevant experimental studies and then calculating the number

of studies which supported or refuted a particular relationship. The limitations of this vote-counting strategy have led to strong criticism. First, simply defining a relationship as significant or nonsignificant gives no information about the magnitude of the effect. This sin of omission is particularly grave in the field of education which tends to ask descriptive questions. Indices of magnitude are especially helpful for educators who are not asking questions that are likely to be clearly black or white (Abrami, Cohen, & d'Appollonia, 1988). A second limitation of vote counting is that if four significant studies are averaged with four nonsignificant studies, the net result is no effect; this result is not descriptive of the original data. A third problem with vote counting is that it yields limited useful information. As a method of research integration it dismisses valuable descriptive information (Glass, McGaw & Smith, 1981).

Shortcomings of procedures for synthesizing research combined with the growing need to apprehend the current and accumulating knowledge on a particular topic, led to the development of methods of research integration which are most commonly referred to as meta-analysis. The purposes of meta-analysis are 1) to describe a body of studies, 2) to summarize the aggregate magnitude of effect of a particular treatment, 3) to identify variables that influence the study outcomes, 4) to recommend directions for future research, 5) to resolve conflicts between opposing theoretical stances, and 6) to make recommendations based on review findings (see Blimling, 1988; Abrami et al., 1988).

Gene Glass pioneered meta-analytic procedures over 20 years ago. In his 1976 speech as president of the American Educational Research Association, the same organization to which Senator Mondale expressed his frustration five years earlier, Glass articulated the need for this groundbreaking research methodology:

> Before what has been found can be used, before it can persuade skeptics, influence policy, affect practice, it must be known. Someone must organize it, extract the message . . . . We face an abundance of information. Our problem is to find the knowledge in the information. We need methods for the orderly summarization of studies so that knowledge can be extracted from the myriad individual researches. (Glass, 1976, p. 4)

This speech marked the birth of a research method which has revolutionized our ability to understand what we know. Today there are over 1000 meta-analyses indexed in ERIC. Particularly in education, where discoveries are fragile and conclusions tend to be in opposition across studies, meta-analysis has proven to be a useful, albeit controversial, tool.

Definition of Meta-analysis

Meta-analysis is the formal, systematic synthesis of the results of individual studies on a particular topic. This process, by Glassian definition, involves integrating the findings of individual studies by statistically analyzing the summary data in each study. Pillemer and Light (1980) accentuate the term "formal" in describing meta-analytic techniques.

> We use the word formal to indicate that they are not specific to a particular reviewer or to a particular set of studies. In fact, their systematic nature is their primary strength. Two reviewers using the same synthesizing procedure should arrive at the same statistical output, although their interpretations of the output may differ. (p. 177)

Advocates of meta-analysis describe the procedure as objective, straightforward, and informative. Meta-analysis offers researchers an additional tool for describing conflicting empirical results about educating children (Brown & Brown, 1987). The

writer uses the comprehensive, quantitative review to translate a body of research into more scientifically accurate and comprehensible generalizations that are likely to have direct implications both for practice and policy (Pflaum, 1982). While fans of meta-analysis are many, criticisms of the method remain healthy (see Eysenck, 1984; Neilsen, 1993; Slavin, 1984).

Criticisms of Meta-analysis

The development and increased use of meta-analysis has turned the research field into a "battle field" (Kraiger, 1985). A number of criticisms have been launched against meta-analysis and those throwing the tomatoes are doing so with force, as reflected in the following statement by Neilsen (1993): "The problems with meta-analysis are legion. I see it as statistical play" (p. 356). She further suggested that taking meta-analysis to the extreme of using it as a vehicle to guide policy decisions or standards for instruction "boggles the mind". The controversy surrounding meta-analysis stems from a number of issues. The primary criticisms of meta-analysis are that meta-analysts are too liberal with their criteria for study inclusion, can introduce bias to the meta-analysis, compare studies that are too different, and are too far removed from the individual subjects of the original studies.

Broad inclusion of studies. First, Glass and his colleagues are accused of giving too much consideration to studies of low-quality. According to Glass (1978), the only difference in identical studies is error. Even studies of the same phenomena are different; the point of interest is how these differences influence outcomes. Slavin (1984) stated that a critical assumption behind meta-analytic procedures is the idea that all studies within the realm of a broad definition, regardless of the quality of the study, should be included in the analysis. Part of the rationale for this assumption is that if reviewers begin applying strict selection criteria, they may be driven by their

own biases to develop criteria that guide them in the selection of studies that support their preferences and in the rejection of studies that oppose them. Slavin (1984, 1986) expressed concern for this procedure and suggested three criteria for including studies: 1) relevance, 2) methodological adequacy in minimizing bias, and 3) external validity.

Kulik and Kulik (1989), on the other hand, argued that Glass and Smith have actually accentuated the need for high-quality studies and the influence that the quality of a study, particularly in terms of bias, can have on a review's results. Kulik and Kulik addressed the issue: "To criticize Glass for paying too little attention to study quality and publication bias is to miss the point of his meta-analytic activities" (p. 232).

Eysenck (1984) expressed strong opinions about the broad inclusion of studies recommended by Glass for meta-analyses.

> some studies are bad in the sense that in their design or analysis they disregard essential parts of the hypothesis to be tested. This means that they ought to be excluded from consideration of the theory in question; their retention on the basis that exclusion is subjective is nothing more or less than an absurdity. (p. 42)

Such an argument advocates good judgment over blind averaging. Eysenck (1984) goes on to detail the specific deficiencies he sees in the meta-analysis of psychotherapy research conducted by Smith and Glass (1977). Eysenck (1984) concluded, "It would be idle to continue listing the sins against inclusivity committed by the proponents of this false god" (p. 56). Such impassioned criticism is indicative of the strength of conviction of researchers on either extreme. What is manna in one camp is blasphemy in another.

Abrami and colleagues (1988) suggested finding a middle ground between the broad inclusion criteria advocated by Glass and the best evidence sampling recommended by Slavin. They noted their own difficulties in judging criteria for inclusion and suggested that such criteria should be thoroughly described in the meta-analysis. In cases where frequently cited studies are excluded, a careful defense should be articulated. Studies that are excluded a priori need to be clearly invalid.

Researcher bias. A second criticism of meta-analysis is that, while specific procedures are defined to make the review process more formal, these procedures provide opportunities for researchers to introduce bias. Cook and Leviton (1980) describe three ways of introducing bias to a quantitative review. First, a narrow literature search can omit relevant data. Secondly, discovered studies may be omitted because of methodological flaws identified by the reviewer. Thirdly, a reviewer may consider the underlying theoretical constructs of a particular study irrelevant to the review.

Comparing apples and oranges. A third criticism of meta-analysis is that it compares apples and oranges. Reference to this fruity analogy is widespread. "Apples and oranges" were addressed, either in efforts to criticize or refute criticism, in approximately one third of the papers on meta-analysis I assembled. Glass (1978) responded to this criticism by saying that the aggregation of apples and oranges is helpful in studying "fruit," and that it is often more informative to investigate broad constructs than subcategories. Kulik and Kulik (1989) further refuted this criticism by explaining that reviews must be adequate in scope to generate meaningful results. Swanson (1996) responded to the mixed fruit metaphor by saying, "Apples are not oranges, but together they both contribute to a good fruit salad" (p. 214). Swanson further explained that while apples and oranges differ, they may weigh the same;

reviews that synthesized the research of studies that are replications of one another would be far too limited. Finally, in a humorous stab at the debate over fruit, Kraiger (1985) wrote, "we seek to tell the apple from the orange. But you try to tell us that all fruit is tasty. We are safe and secure in our confusion. Leave us be" (p. 800). The persistence of the apple-orange analogy indicates the relentlessness of opponents and proponents of meta-analysis.

Implications of review findings. A final criticism of meta-analysis, particularly in the educational arena, is that studies of studies are too far removed from actual subjects. Neilsen (1993), speaking to the use of meta-analysis in education, drew an analogy between meta-analysis and the story circle game in which the first person whispers a sentence to the second person, the second to the third, and so on. The sentence that emerges from the final person in the circle bears little resemblance to the original. Neilsen also reminds us of Parker's comment that research without a face is without a point. She goes on to explain that our confidence in the meta-analytic process hinges on our belief that we can effectively present what was investigated originally, that the original investigation accurately represented a particular phenomena, that varying levels of bias in the original studies can be statistically washed away, and that the developing quantitative review, which is a composite of deferred meaning, can produce guidance for future research.

Neilsen (1993) argued, "Meta-analysis, regardless of what claims to truth, understanding, or direction we hold for it, drives us further inside a vortex of self-referential discourse and away from the voices of teachers and children in the field" (p. 351). Neilsen continued by comparing research to a food chain, explaining that the higher we go in the research chain the farther we are removed from actual

children and schools and the more we perpetuate the idea that most authentic, reliable, and critical research is formulated outside the classroom.

Pflaum (1982) holds an opposite view. She maintained that, ideally, it is the accumulated knowledge from a body of studies which should direct classroom practice. Pflaum further stated that the technique of meta-analysis allows researchers to more effectively manage the translation difficulties associated with getting research into the classroom. Meta-analysis increases a researcher's ability to identify reliable patterns in educational research and communicate these findings to practitioners in the field.

Baldwin and Vaughn (1993) launched a criticism similar to that of Neilsen (1993) by asking whether the meaning of a group of studies can be derived from the sum of their independent conclusions. They further questioned whether the quality of a body of research can be deduced from the sum of their methodological shortcomings. In a criticism of a meta-analysis of secondary reading research, Baldwin and Vaughn stated: "What the authors have failed to recognize is that empirical efforts do not add up to truth; they converge on it over time" (p. 355). In general, critics along this vein maintain that summaries of what is perceived as truth may in fact contain less truth in combination that they did when evaluated on individual merit.

Conclusion. In summary, meta-analysis has both vocal opponents and vocal proponents. Criticisms that the method is too liberal or too conservative abound. Many of these arguments have merit; meta-analysts face challenges in the conception and in the execution of a review. One of the strongest faults articulated by critics is that, while meta-analysis requires reviewers to make fewer judgments and is thus more formal and objective, the process is still fraught with decisions about the

theoretical relevancy of studies, the qualifying methodological considerations, and the actual meta-analytic technique employed (Cook & Leviton, 1980). Cook and Leviton ended their defense of traditional, qualitative reviews by stating that it was not accidental that their discussion made the best of qualitative and quantitative reviews look similar. They intentionally tried to demonstrate that the frequently cited flaws of qualitative reviews are not inherent in the method and that the strengths of meta-analysis can be used to improve upon narrative reviews. Cooper and Arkin (1981) suggested that literature reviews fall on a continuum from exclusively qualitative to purely quantitative. For example, Chall's book, Learning to Read: The Great Debate (1983), contains a review of the literature dealing with phonics instruction. Chall coded studies, examined statistics, and interviewed subjects, resulting in a study with strong quantitative and qualitative elements. While the line between qualitative reviews and quantitative reviews may not be black and white, meta-analysis has an advantage over traditional reviews in that the former has clearer criteria for judging its own quality (Abrami, et al., 1988).

Steps in Conducting a Meta-analysis

Meta-analysis is beginning to share equal status with primary research studies. However, because the original research in a meta-analysis is actually an aggregation of research previously conducted, meta-analysts have to deal with the perception that conducting quantitative reviews is not as scientifically exacting as conducting studies involving primary research. Carlberg and Walberg (1984) spoke to the challenge of assimilating research evidence: "The logic that underlies quantitative research synthesis is simple, compelling, and elegant. It is also, therefore, seductive; in truth, a credible, rigorous synthesis is usually much more complex an undertaking than is a primary study itself" (p. 25). Thus, while meta-analysis can be broken into distinct

steps which can be clearly articulated, conducting such a quantitative review is likely to be a very demanding task which, like primary research, involves making judgments specific to the study.

A number of researchers present steps for conducting a meta-analysis (Abrami et al., 1988; Jackson, 1980; Johnson, 1989; Kulik & Kulik, 1989). The process is similarly described by most. The first step is specifying the question being investigated. This step also involves defining terms relevant to the study and detailing the relationship to be examined. Second, a meta-analyst completes an exhaustive search of the literature for pertinent studies. Third, the synthesist codes articles according to characteristics that are general to most studies and characteristics that are specific to the particular collection of studies. The fourth step in conducting a meta-analysis is statistically synthesizing the individual study results. Finally, a meta-analyst must interpret the results.

Defining the Question

The first step in conducting a meta-analysis is defining the area under investigation. With as much theoretical and methodological clarity as possible, researchers should define which variables, under which conditions will be examined in the review. This process involves articulating the X and Y variables as well as W variables, or moderators. Moderator variables are those which can be expected to alter the magnitude and/or direction of the relationship between X and Y (Johnson, 1989).

Meta-analyses often deal with broad, loosely defined areas such as, psychotherapy, whole language, or mastery learning. Such terms may mean one thing to one person and something very different to someone else. Whole language, for example, may mean anything from using children's literature for instructional

purposes to not explicitly teaching reading at all. Thus, while defining the X and Y variables may be a difficult task, it is critical to developing a strong meta-analysis.

These definitions of variables drive literature searches and should be broad enough to encompass enough studies and specific enough to establish a conceptual framework. Extremely narrow definitions provide minimal information about whether a finding is applicable in various situations (Cooper, 1982). Consequently, "meta-analysts who employ broad conceptual definitions can potentially reach more definitive and robust conclusions than reviewers using narrow definitions" (p. 294). The particular capability of meta-analysis to deal with large numbers of studies gives reviewers the opportunity to work with broad conceptual definitions (Cooper & Arkin, 1981). Carlberg and Walberg (1984) explained that the focus of a meta-analysis falls somewhere on a continuum between being so narrow in scope that relevant treatment variations are lost, and being so broad in scope that marginally relevant research is included. Blimling (1988) suggested that, while definitions of experimental variables should be established in advance of the meta-analytic process, these definitions may be refined throughout the procedure.

Searching the Literature

After defining the area of research, a meta-analyst must obtain studies. Literature searches should be comprehensive, as the goal of an integrative review is to encapsulate the accumulated state of knowledge regarding the relationship under investigation and to highlight important questions that research has left unanswered (Cooper, 1982). Searching the literature is as critical to the meta-analytic process as random selection is to primary research (Blimling, 1988).

While most meta-analysts by definition will conduct broad searches of the literature, researchers are not necessarily equal opportunity reviewers; all studies do

not have the same odds of being retrieved by a reviewer. Unpublished studies have the least opportunity for discovery. Studies that contradict the public perception tend to be less accessible and variability in review techniques will lead to variability in conclusions (Cooper, 1982).

There are at least five methods for retrieving data (Johnson, 1989). In the ancestry approach researchers examine the reference list of previous studies and narrative reviews. The descendancy approach involves locating a critical article and tracing it to other studies that have cited it. Computer databases can be employed to locate abstracts using keywords. With the invisible college technique, meta-analysts utilize a network of researchers who have a finger on the pulse of the research community. In this way, reviewers can locate "fugitive literature" which may not have been published. Finally, meta-analysts can make manual searches of relevant journals. This final step is most logically conducted with most recent journal issues as those are not included in data bases. Abrami and colleagues (1988) also suggested that to increase the odds of locating unpublished research, meta-analysts should write letters to active researchers in the field of interest and scan programs of relevant professional conventions.

In order for a completed meta-analysis to stand up under scrutiny, careful and thorough search techniques must initiate the project. A primary goal for reviewers is to protect the validity of their reviews by accessing as many sources of information as possible.

> The quality in this step is essential to the quality of the meta-analysis, because it takes the place of the careful sample design that is characteristic of primary research. If a meta-analysis is based on a handy collection of studies without a

> serious search of all available materials, it has lost its claim to validity. (Manke, 1988, p. 550)

While this process bears little difference from the preliminary searches guiding narrative reviews, in a meta-analysis the exacting process of tracking relevant studies requires more attention and care than that in a traditional literature review (Carlberg & Walberg, 1984).

Even the most ambitious and sophisticated researcher is unlikely to locate all of the primary research on a given topic. To avoid critics charging bias in study selection, quantitative reviewers should include comprehensive descriptions of their search procedures. This will allow those judging the final product an opportunity to evaluate the representativeness of the base of data (Kavale, 1988).

Coding Studies and Determining Criteria for Inclusion

Once the studies have been identified, meta-analysts must make decisions in regard to screening procedures. Glass, McGaw, and Smith (1981) advocate broad acceptance of studies regardless of possible flaws in methodology. Slavin (1986), on the other hand supports a "best-evidence synthesis". Both views have received criticism. Eysenck (1978) stated that Glass' use of all relevant studies is a case of "garbage in and garbage out" (p. 517). Kulik and Kulik (1989), on the other hand, have found the best-evidence synthesis prone to bias because criteria for determining which is the best-evidence are ill-defined.

Cooper (1982) stated, "The use of any evaluative criteria other than substantive methodological discriminations is a threat to the validity of a research review" (p. 297). While Cooper did not suggest specific criteria for reviewers to evaluate research, he did suggest that reviewers should develop their criteria for evaluation before the literature is searched. Prior to literature retrieval, exclusion

criteria should be explicitly and objectively stated. Chall (1983) described her inclusion criteria for Learning to read: The great debate and included her "Schedule for Analyzing Studies". She commented:

> Although most studies were unsatisfactory in some respects, I assumed that all the authors were honest researchers searching for honest answers, and I looked for the grains of underlying truth to be found in each study. Had I considered only studies that fulfilled all necessary experimental conditions, I would have been left with just a handful--if that many. (p. 102)

Pillemer (1984) cautioned that it is critical to distinguish between methodological rigor and rigidity. While meta-analysis has the laudable benefit of a heightened level of objectivity, the goal of meta-analysis should be increased flexibility of method rather than restrictive uniformity.

The purpose of the coding process is to describe the various characteristics of each study. The process forms a statistical portrait of features of the different studies. When the analyst is interested in examining the relationship between study features and effect size, only variables with adequate variation need be examined. However, as meta-analysis should help us determine which settings have been over or understudied, reviewers may code studies for features that do not vary. For example, a reviewer may code the intelligence quotient of subjects in order to compare subjects of high, average, or low intelligence. If coding reveals that only subjects of average intelligence have been studied, then intelligence quotient cannot be explored as a moderator variable. However, the coding remains meaningful in that it revealed that future research should include subjects of high or low intelligence.

Johnson (1989) offered two recommendations in deciding what features to study. First, reviewers should examine previously conducted meta-analyses to see

which characteristics have been coded in the past. Using similar coding strategies between meta-analyses helps ensure that the reviews are cumulative. Secondly, reviewers should read a sample of the studies to be coded. An intimate understanding of the research in an area is necessary to avoid missing an important feature of a body of research. Coding categories will include characteristics that are unique to a particular body of research: design features, setting features, and publication features.

Kulik and Kulik (1989) stated that most meta-analysts have not paid enough attention to the reliability of their coding procedures. Glass and his colleagues (1981) found adequate reliability on the type of coding that is typically done by meta-analysts. Other examinations of coding reliability are sparse, although this is an important issue. Johnson (1989) suggested that the only way to know whether the results of a meta-analysis are valid is to have at least two independent raters conduct the coding. Hall, Tickle-Degnen, Rosenthal, and Mosteller (1994) describe coding variables as low-inference or high-inference and suggest that coding forms that investigate characteristics which are highly subjective, such as study quality, use some form of interrater reliability.

Kulik and Kulik (1989) offered an alternative to reporting inter-coder reliability coefficients. They suggest that reviewers include in their reports the main features of all coded studies. If this detail is included, other researchers can make their own determinations as to whether studies have been appropriately coded. A further benefit of including specific features coded is that it allows future researchers to expand on a particular meta-analysis making the process cumulative.

The primary goal of the meta-analyst is to integrate as much relevant data as possible. The steps of defining variables, locating studies, and coding studies can all influence this goal. "These features should be developed and explained with the same

care as if they were hypotheses from primary research" (Abrami et al., 1988, p. 166). Thus, a meta-analyst should be meticulous in describing and implementing these critical steps.

Conducting Statistical Analysis

Meta-analysis is most appropriately viewed as a flexible addition to a researcher's storehouse of analytic procedures, with its use guided by project-specific goals rather than a canon (Pillemer, 1984). Blimling (1988) explained, "The statistical analysis of meta-analytic data is a method for describing a complex data set comprised of study outcomes for the purpose of making this information more accessible and understandable" (p. 546).

Calculating effect sizes. There are a number of ways to statistically synthesize a collection of related studies. In this study effect sizes will be used. Kulik and Kulik (1989) defined effect size as "a general measure of the magnitude of a treatment effect on a dependent variable, expressed in such a way that the treatments in many different studies can be directly compared" (p. 263). Cohen (1977) explained that effect sizes represent "the degree to which the phenomenon is present in the population, or the degree to which the null hypothesis is false" (pp. 9-10). Effect sizes are expressed in standard deviations and, because they are free of the original unit of measurement, they can be compared and combined with effect sizes from other studies examining the same relationship. The most commonly utilized effect size is Cohen's d which is computed by taking the difference in the measure of the dependent variable in the mean of the control group and the mean of the experimental group and dividing by the pooled standard deviation of the control group and the experimental group (see Blimling, 1988).

Effect sizes typically calculated in meta-analyses are biased estimates of the underlying effect. Hedges presented one formula for correcting this bias. .Kulik and Kulik (1989) calculated 27 effect sizes with and without the Hedges correction. They found that the two effect sizes correlated .999 and that, in "most cases", they agreed to the second decimal place. Thus, many researchers do not correct for this bias. However, differences in the biased and unbiased estimates will be larger for studies with smaller sample sizes (Johnson, 1989).

Combining effect sizes. Once effect sizes have been calculated, they must be combined. Given that studies are not perfect replicates of one another, researchers have explored various methods for weighting studies. Hedges recommends computing a composite effect size by calculating an average d value, the d+, with each effect size weighted by the reciprocal of its variance. Using this procedure, studies with the largest sample size, and presumably the most reliably estimated outcomes, are given the greatest weight (Hedges & Olkin, 1985). Shadish and Haddock (1994) stated that the Hedges procedure is by far "the most widely accepted weighting scheme" (p. 264).

Typically, a 95% confidence interval is drawn around the average effect size. This means that a researcher can say with 95% confidence that the real population effect size is within the given range. If the value zero (0.00), which indicates no difference, is included in the confidence interval, then it may be concluded that across all studies there is no relationship between the independent and dependent variables (Johnson, 1989).

As the mean weighted effect size is computed, analysts should examine the homogeneity of the d values in order to determine whether the combined studies can be adequately described by a mean effect size (Hedges, 1981; Hedges & Olkin,

1985). If the effect sizes are homogeneous, not varying dramatically in magnitude or direction, then the analyst can conclude that the mean effect size is representative of the sample of studies. If on the other hand, the effect sizes are heterogeneous, a researcher must explore the moderator variables which influence the effect sizes.

Model testing. Moderator variables are those variables which can influence the magnitude of the effect. Moderator variables are typically identified for examination in the coding stages of the meta-analysis. Such variables may include gender, age, time of training, setting, etc. Meta-analysts must be selective in identifying moderators to investigate, seeking a balance between examining relevant connections and exploring all significant relationships. "Needless to say, the choice and definition of moderators can profoundly affect the conclusions of a research synthesis" (Hall, Tickle-Degnen, Rosenthal, Mosteller, 1994, p. 26).

If effect sizes from a research synthesis are found to be heterogeneous, they are commonly sorted into separate groups according to particular characteristics of the primary studies. This categorical model, defined by Hedges and Olkin (1985), is comparable to an analysis of variance. The categorical model can be tested between-groups of studies, similar to testing for a main-effect, and within groups of studies. Hedges (1994a) describes this procedure as "exploring the relationship between a categorical independent variable (the grouping variable) and effect size" (p. 286). Hedges also offers a continuous model for examining the heterogeneity of effect sizes. His procedures are analogous to conventional regression analysis. In continuous model testing, a researcher attempts to explain the variability in the effect sizes by using continuous quantitative study attributes as predictors (Johnson, 1989).

As previously discussed, meta-analysts statistically summarize the data collected from other researchers. However, by identifying moderator variables

between studies, researchers can test new hypotheses across studies. In these cases, the moderator variable is usually a characteristic of an entire research report. The results of these original analyses are referred to as "review-generated evidence". Eagly and Wood (1994) maintained that review-generated results have limited construct validity because subjects in the original research were not randomly assigned to levels of the moderator.

Interpreting Results and Drawing Conclusions

As previously discussed, the statistical procedures in a meta-analysis have aroused both enthusiasm and distrust and have led to a considerable professional interchange (Pillemer, 1984). Enthusiasts see meta-analysis as providing liberation from the shaky, traditional literature review which has been disappointing in its ability to influence policy or help us draw conclusions (Glass, 1978; Kulik & Kulik, 1989; Pflaum, 1982). Critics argue that meta-analysis oversimplifies issues, glorifies the effect size, and drives researchers away from the subjects who may eventually be influenced by the conclusions (Baldwin & Vaughn, 1993; Eysenck, 1984; Neilsen, 1993).

In addition, interpretation of effect sizes has caused some debate. While researchers agree that effect sizes are valuable indices, they do not agree on their meaning (Cooper, 1981). Rosenthal and Rubin (1982) found that even skilled statisticians and skilled researchers were unable to intuitively interpret the relevance of an effect size. Cohen (1988) has offered some guidance by categorically defining effect sizes. By his definition, a small effect size is .20; a moderate effect size is .50 and a large effect size is .80. Cohen established these definitions based on his informal examinations of the magnitude of effects typically yielded by psychological research.

Even considering its limitations, explanations of variations in study findings can be understood by meta-analysis. Cooper and Arkin (1981) explained that the one thing that can be assumed about the effect size estimates of a particular meta-analysis is that they are descriptive of the particular collection of studies. Given such limits, effect size is a legitimate statistic. The unbiased inference that can be made from the effect size of one meta-analysis is that other literature reviews, utilizing similar retrieval processes, would arrive at similar results.

Threats to validity. While meta-analysis has been generally accepted as a sound option for integrating studies, a number of threats to the validity of a research synthesis exist and need to be explored in the final sections of the meta-analysis. These include but are not limited to missing data from primary studies, violation of the assumption of independence, failure to appropriately weight studies, publication bias, and lack of statistical power (Matt & Cook, 1994). These threats are discussed in greater detail below.

The first threat to the validity of a meta-analysis involves unreported data in primary studies. Significance tests with nonsignificant results are sometimes unreported. Researchers may state that a test was not significant but fail to present the corresponding statistics. While researchers may not report the numerical results of statistical tests, they typically present the means and standard deviations for their dependent variables. Meta-analysts may use the original means and standard deviations presented in a study to calculated effect sizes and circumvent the problem of unreported figures.

Other types of unreported data may lead to inaccurate coding. Meta-analysts are limited by what is presented in the primary study; primary researchers are limited by publication space. If the primary report lacks detailed descriptions of assignment

to groups, training, and evaluation, meta-analysts may make inaccurate assumptions in the process of coding studies.

Second, in conducting a meta-analysis one must consider the assumption of effect size independence. That is, subjects from a particular study should only be represented once in the aggregation of effect sizes. Hedges (1994b) describes four particular violations of independence which are common in meta-analyses. First, if the same subjects are used when effect sizes are calculated with different measures, it is a violation of the assumption of independence of effect sizes. Second, different experimental groups compared to the same control group are a violation. Individual studies represented more than once in a meta-analysis are the third common violation. Finally, if a series of studies conducted by the same research team is included in the meta-analysis, the assumption of effect size independence is violated.

One standard way to circumvent the problem of effect size dependence is to use a mean effect size from each study which offers more than one relevant effect. However, given an interest in investigating particular variables that differ within a study, multiple representations of the same study can be justified. Johnson and Eagly (in press) makes this point when he states, "Despite these concerns, multiply representing studies may be defensible to address certain meta-analytic questions" (p. 55). Thus, a researcher should look for a balance between pursuing answers to interesting questions and over-representing some studies.

A third threat to the validity of a meta-analysis is failure to appropriately weight studies. The goal of a meta-analyst should be to give more influence to studies that are more methodologically sound. The most commonly used weighting method is Hedges and Olkin's (1985) procedure for weighting studies by the inverse of their sample sizes. This weighting procedure is based on the assumption that

studies with large sample sizes are more valid than those with small sample sizes. The Hedges and Olkin weighting procedure is discussed in the previous section on combining effect sizes.

A fourth validity threat is publication bias. Publication bias can be introduced to a study through reporting bias, misrepresentation of the results of primary studies, or through failure to retrieve all relevant research (Greenhouse & Iyengar, 1994). Publication "bias occurs when effect estimates are selected that are just as substantively relevant as those not selected but that differ from them in average effect size" (Matt & Cook, 1994, p. 508). In other words, bias is introduced to a meta-analysis when omitted studies have effect sizes that tend in the opposite direction of included studies.

A number of researchers have developed analytic techniques for dealing with publication bias in a meta-analysis. For example, Light and Pillemer (1994) presented a graphic method for the detection of publication bias. They graphed sample sizes and effect sizes to create "funnel plots" which illustrate the degree of bias in a meta-analysis.

Other methods for detecting publication bias exist, but none, including the one just described, are commonly used. Cooper and Hedges (1994) stated, "While these emerging methods are promising, even their most enthusiastic advocates suggest that it is better not to have to correct for publication bias in the first place" (pp. 425-426). The most important step in preventing publication bias is aggressively seeking out unpublished studies.

Finally, validity is threatened in a meta-analysis if analyses lack statistical power. Because a the sample size for a meta-analysis is the sum of the sample sizes of studies within the synthesis, the statistical power of a meta-analysis is typically

high in comparison to that of primary research. However, if an analyst conducts model testing and analyzes the relationships between subgroups within variables in the body of studies, where subgroups are small, power is reduced. For example, if a researcher is examining gender as a moderator variable, the analysis of subgroups (male and female) will have low power if one of the subgroups has only a few cases (i.e., 12 males vs. 2 females).

Conclusion

This section has served as a blueprint for conducting a meta-analysis. The discussion presented commonly used procedures which are relevant to this particular meta-analysis. These steps to conducting a meta-analysis should be described in such detail that another researcher could replicate the study exactly (Abrami et. al, 1988). These steps include defining the relationship to be explored, conducing a comprehensive search of the literature, coding studies according to characteristics that will later be analyzed, synthesizing studies statistically, and drawing conclusions and interpretations from the results.

Examples of Quantitative Reviews

Examples of meta-analyses are a logical extension of the previous presentation of steps in conducting a quantitative review. The following summary of reading research indicates the versatility of meta-analysis and offers examples of the types of conclusions meta-analysts reach.

Comprehension

Haller, Child, and Walberg (1988) conducted a meta-analysis of the research on teaching metacognitive strategies in reading. Their study, "Can Comprehension be Taught? A Quantitative Synthesis of 'Metacognitive' Studies," compiled 115 effect sizes from 20 studies. The mean effect size was .71, "which indicates a substantial

effect" (p.5). The authors provided a detailed description of their search and coding procedures. They pointed out that their overall effect size was one of the largest reported in a meta-analysis in the field of education at that time. Haller et al. concluded that comprehension can be taught and that, while it is helpful in all grades, it is particularly effective in the seventh and eighth grades.

Rosenshine and Meister (1994) investigated reciprocal teaching and its effects on comprehension. The authors examined 16 published and unpublished studies. They coded studies both on generic and content specific features and classified studies as high, medium, or low in quality. Detailed descriptions of all these procedures are included in their article, "Reciprocal Teaching: A Review of the Research." Rosenshine and Meister analyzed their effect sizes across a number of features of the studies including but not limited to, the grade level of the students, the number of instructional sessions, the size of the instructional group, the number of strategies taught, the type of control group, and whether the group was taught by the experimenter or the teacher. Overall, when experimenter-designed tests were used, students in the reciprocal teaching group had significantly higher scores than the students in the control group, with a strong median d of .88. The difference in the two groups was also significant when standardized tests were used; the median d was weak, .32.

"The Effects of Vocabulary Instruction: A Model-Based Meta-analysis," a study conducted by Stahl and Fairbanks (1986), investigated two questions: Does vocabulary instruction have a significant effect on children's comprehension of text? What types of vocabulary instruction are most effective? The researchers used published studies that were referenced in ERIC or located by cross referencing bibliographies. They did not include unpublished studies. Stahl and Fairbanks

attributed a strong mean effect size of .97 to vocabulary instruction for comprehension of passages containing previously taught words. The analysis suggested that the most effective methods for teaching vocabulary included providing both contextual and definitional information, giving students several exposures to the word being studied, and engaging the students in deeper processing. Stahl and Fairbanks offered other effect sizes and conclusions specific to the duration of the instruction, the size of the group, the type of control group, and the method of assessing knowledge of words.

In a study examining the relationship between adjunct pictures in text and comprehension of the text, Readence and Moore (1981), aggregated the results of 16 studies. Their article, "A Meta-analytic Review of the Effect of Adjunct Pictures on Reading Comprehension," offered a sketchy description of their search and coding procedures. Furthermore, half of the studies they included were by the same author. In addition, they included studies for which the results were unclear. An example of such inclusion is in the following statement:

> It should be pointed out with regard to the overall results that in one study (Dwyer, 1968), the exact direction of six effects was not clearly specified. For the purpose of this analysis, these effects were interpreted to be in favor of the adjunct picture group. (p. 220)

Readance and Moore concluded there is a small effect of adjunct pictures on reading comprehension. On average, approximately 5% of the variability in learning from text can be explained by knowing if the subjects looked at the pictures. The authors further concluded that there was a large effect for university-level subjects in traditional text settings.

Grouping for Instruction

Slavin's best-evidence synthesis was mentioned in the previous section describing coding procedures and study inclusion criteria. Slavin (1987) conducted a best-evidence synthesis which compared different models for grouping students for instruction. In this study, Slavin provided a description of his inclusion criteria. He also provided detailed descriptions of the individual studies. Slavin chose to report the median as the measure of central tendency because the median is not effected by outliers. Slavin arrived at a median effect size of .00 for assignment of students to self-contained classrooms by ability. He also found a moderate median effect size (.45) for the Joplin Plan, which regroups students across grade-levels for reading instruction. Within-class grouping was only found to be instructionally effective for mathematics, with a median effect size of .34. Critics have argued that Slavin introduces bias to his best-evidence syntheses by excluding relevant studies (see Kulik & Kulik, 1989). Another possible bias is that the Joplin Plan is the model used for Slavin's reading intervention, Success For All, which was founded in 1986, one year before this study was published.

Peer Tutoring

Mathes and Fuchs (1991) also conducted a best-evidence synthesis. They examined the efficacy of peer tutoring in reading for students with disabilities. They located 30 studies, but only 11 met a priori standards for inclusion in the analysis. The average, unbiased effect size was moderate (.40) for studies comparing peer-tutoring to groups receiving reading instruction with no intervention. The average effect size was weak (.14) for studies comparing peer-tutoring to groups which received a control intervention guided by the researcher. Mathes and Fuchs concluded that peer tutoring with disabled students has a greater effect on reading achievement than typical classroom instruction and that it is equally effective as other

researcher-led interventions implemented by the classroom teacher. The authors also examined effect sizes across specific characteristics of the studies, such as academic knowledge of the tutor, focus of the tutoring session, setting of tutoring, and the role students play during tutoring.

Learning Styles

Davies (1995) investigated the effectiveness of matching instructional methods or testing conditions to learning style/reading style preferences. She aggregated the results of 19 studies, both published and unpublished, which yielded 68 effect sizes. She found a weak median effect size of .10 when learning/reading styles were matched for instruction or testing. Davies explained that the median was the most appropriate measure of central tendency because the effect sizes were skewed. These results varied slightly for measures of reading comprehension and measures of reading rate and accuracy. Interestingly, Davies also found a statistically significant difference between the mean effect sizes of studies completed at St. John's University, where the Reading Styles Inventory (Dunn & Dunn, 1975) was created, and those completed elsewhere.

Whole Language

Three meta-analyses comparing the effectiveness of whole language and traditional basal instruction exist in the literature. Stahl and Miller (1989) conducted a study entitled "Whole Language and Language Experience Approaches for Beginning Reading: A Quantitative Research Synthesis". They calculated 121 effect sizes from studies in which the experimental group 1) emphasized using the child's own language as a medium of instruction, 2) was child-centered, 3) emphasized trade books over basals, and 4) taught decoding as needed in the context of reading whole texts. The results of both vote-counting and statistical summary suggested that the

two methods are equally effective but with different subsets of the population. Whole language/language experience approaches were more effective than traditional instruction with kindergartners than with first-graders. Whole language/language experience approaches were less effective than traditional approaches with students who were specifically labeled disadvantaged.

In 1993 Stahl revisited the whole language vs. traditional instruction issue. Stahl, McKenna, and Pagnucco (1993) reviewed the effects of whole language instruction using both vote-counting and meta-analysis procedures. They analyzed 14 studies, which were located through ERIC, PSYCHLIT, and Comprehensive Dissertation Abstracts databases and were published over a five year period, 1988-1993. Several factors made the data difficult to aggregate. First, research on whole language tended to use affective measures rather than measures of reading achievement. Secondly, measures of reading comprehension varied as some used standardized tests and some used comprehension questions from an informal reading inventory.

Stahl and his colleagues concluded that whole language approaches seemed to be more effective in kindergarten than when compared to formal reading instruction in first-grade. They also concluded that "whole language does not appear to have the effects on attitude that are claimed" (p. 9). Stahl and associates further examined research on Reading Recovery, literature discussion groups, and supplemental literature programs. They closed by advocating balance in instruction:

> The apparent contradiction is that many practices arising from whole language are highly effective as are many arising from traditional practice. Practices drawn from both are needed to meet the different needs of children. The contradictions come from the nature of children learning to read, not from an

externally imposed mandate....When good teachers try to meet students' needs, what they will do will usually transcend philosophy and politics. (p. 21)

Gee (1995) also examined the effectiveness of whole language instruction, particularly in relation to reading comprehension. He statistically summarized 21 published studies with a resulting average effect size (.65) moderately favored whole language instruction. The author examined specific variables in studies such as the length of the intervention, the ability of students, the design of the studies, the quality of the studies, and size and characteristics of the samples. Gee concluded, "Nearly every study analyzed showed a positive effect size in the direction of the whole language approach" (p. 15). He also noted that effect sizes were even stronger for studies using experimental rather than quasi-experimental designs. However, sample size was negatively correlated with effect size; smaller groups had larger effects.

The meta-analyses on whole language were the only instance where more than one meta-analysis on a particular topic was encountered. The somewhat contradictory findings seem to compromise the idea that a meta-analysis can help researchers arrive at solid conclusions. However, the reviewers did not aggregate the same body of studies, nor did they ask the same questions. Even more relevant, some of the inconsistency is probably due to the vagueness of the construct. Hall, Tickle-Degnen, Rosenthal, and Mosteller (1994) stated, "Two synthesists may not define the domain in the same way even though they want to study the same fundamental hypothesis" (p. 18). Defining "whole-language" and "traditional-instruction" across studies promises to be challenging at best and inconsistent at worst. If these researchers asked the same questions in an analysis of the same set of experiments, it would be a truer test of the meta-analytic method.

Conclusion

Reading research has expanded rapidly and the need to capture what investigators have learned has become more pressing. This abundance of information has encouraged reading researchers to embrace meta-analytic techniques. The conclusions drawn deal with topics ranging from comprehension to grouping for instruction.

Conclusion

This chapter pursued three avenues of discussion. First, it described the history, definitions, and criticisms of meta-analysis. Second, it detailed the steps in conducting a meta-analysis. The chapter ended with a summary of sample meta-analyses conducted with reading research.

CHAPTER III

Methods and Procedures

Defining the Research Questions

The purpose of this study was to determine the strength of the relationship between phonemic awareness training and growth in phonemic awareness. Through meta-analytic techniques, the following research questions were explored.

1. Does phonemic awareness training impact phonemic awareness?
2. Which, if any, categorical training variables affect phonemic awareness?

- Approach of Training (segmenting vs. blending vs. segmenting and blending vs. other combinations)
- Goal of Instruction (mastery oriented vs. lesson oriented)
- Context of Instruction (supplementary letter names incorporated vs. no supplementary letter names)
- Frequency of Training (1-5 times per week)
- Grade of Subjects at Training (pre-K vs. kindergarten vs. first grade)
- Outcome Measures (segmenting vs. blending vs. deletion vs. combination measures)

3. Which, if any, continuous training variables affect phonemic awareness?

- Quality Rating
- Invalidity Index
- Minutes per Training Session
- Total Minutes in Training
- Size of the Training Group

Searching the Literature

A thorough search for relevant studies both published and unpublished was conducted. In order to maximize the number of primary studies accessed, several search methods were employed. First, relevant databases were searched. These were Educational Resources Information Center (1966-June, 1997) and Dissertation Abstracts International Computer File (1871-September 1997). The databases were searched for the following combinations of descriptors which identified the corresponding number of documents: phonological awareness, 254 documents; phonemic awareness, 161 documents; auditory discrimination and reading, 372 documents. A total of 787 documents (including redundancies) were identified through database searches. The abstracts of these studies were read and all of the studies which dealt with the relationship under investigation, even in the broadest sense, were marked for further investigation.

Through reading the reference lists of articles, 57 additional studies were identified. Finally, a manual search of journals from June, 1997, the final date on the ERIC database, through November, 1997 was conducted. Journals searched manually were Cognition, Journal of Educational Psychology, Reading Research Quarterly, Journal of Literacy, Scientific Studies of Reading, and Journal of Experimental Child Psychology.

Some searching techniques were omitted, such as surveying programs from professional conferences, writing professionals who have conducted related research, and advertising in widely-read journals or on the Internet. However, White (1994) explained, "The point is not to track down every paper that is somehow related to the topic....The point is to avoid missing a useful paper that lies outside one's regular purview" (p. 44). In terms of extensiveness, this meta-analysis includes a comprehensive collection of

applicable published studies and a representative collection of applicable unpublished studies.

Determining Criteria for Inclusion

Of the studies located, 151 which involved training in awareness of the phonemic nature of speech qualified for possible inclusion in this meta-analysis. These studies were further narrowed according to four criteria. First, studies had to have both an experimental and a control group or more than one experimental group. Second, studies had to define phonemic awareness as the conscious manipulation of phonemes and reflect this definition in their training. Third, the phonemic awareness training in the studies had to be specific. Fourth, studies had to deal with prekindergarten, kindergarten, or first-grade children who did not qualify for special education. Finally, the training in the study had to be conducted in English. These narrowing procedures are more thoroughly defended in the following paragraphs.

Control Group

First, studies were eliminated if they did not have a comparison group. All studies which employed a pretest-posttest design with only one group were removed from the sample on the basis of methodological inadequacy. While Glass endorses the broad inclusion of studies regardless of methodological flaws (see Glass, 1976, 1978), other meta-analysts recommend eliminating studies of poor quality (Abrami et al., 1988; Eysenck, 1984; Slavin, 1984). The minimum expectation for quality in this study was use of a control group. Shavelson (1988) addresses the importance of using a control group.

> The control group is very important in dealing with threats to internal validity. Since, for example, control and experimental groups can be treated in exactly the same way, both groups experience the same internal and external history. If the

> experimental group performs better than the control group, history cannot be used to explain the difference. (p. 24)

In addition, if studies without comparison groups were included in this study, they would have to be considered as their own body of research for separate meta-analysis. Thus, the nonexperimental designs were not included.

Phonemic Emphasis

Second, a definitional issue that arose in narrowing the studies to be included in the meta-analysis related to phonemic awareness as it deals with spoken rather than printed language. In some studies, the training emphasized teaching children letter/sound correspondences and then testing their phonemic awareness. For purposes of this study, the earliest and strongest emphasis in the training had to be on manipulating the speech stream with a later and secondary emphasis on attaching graphemes to the sounds with which children were already working. For this reason, studies such as, "How Letter-sound Instruction Mediates Progress in First-grade Reading and Spelling" (Foorman, Francis, Novy, and Liberman, 1991) and "The Effects of a Phonics-oriented Kindergarten Program on Auditory Discrimination and Reading Readiness" (Harckham & Hagen, 1970) were excluded.

Specific Training

Third, studies with vague or general training procedures were eliminated. In order for a study to be included, the experiment had to explicitly or implicitly train children in phonemic awareness. Studies that simply placed children in a particular environment to see if it would effect phonemic awareness were not included. For example, "The Effects of a Code Emphasis Approach and a Whole Language Approach Upon Emergent Literacy of Kindergartners" (Ribowsky, 1985), "A Whole Language and Traditional Instruction

Comparison: Overall Effectiveness and Development of the Alphabetic Principle" (Klesius, Griffith, Zielonka, 1991), and "Developing Phoneme Awareness Through Alphabet Books" (Murray, 1996) were excluded.

Age/grade of Subjects

Next, studies that dealt with subjects in upper-elementary school or high school and adults did not qualify for inclusion. Because phonemic awareness training is associated with beginning reading, educators and researchers typically emphasize it in preschool, kindergarten, and first-grade; unless someone has had difficulty learning to read. In addition, any studies using subjects from special education classrooms were eliminated on the grounds that these studies should be combined for their own meta-analysis. O'Conner and Notari-Syverson (1995) wrote, "There is reason to suppose that treatment effects may differ between typical and very hard to teach youngsters" (p.4). Consequently, in order to avoid aggregating weak effects that were more a representation of the learner than of the training, studies were limited to "typical" students. Studies such as "Phonological Awareness Training and Remediation of Analytic Decoding Deficits in a Group of Severe Dyslexics" (Alexander, Anderson, Heilman, Voeller, & Torgesen, 1991) and "Assessment and Remediation of a Phonemic Discrimination Deficit in Reading Disabled Second and Fourth Graders" (Hurford & Sanders, 1990) were eliminated.

Training Conducted in English

Finally, studies were eliminated from inclusion if the training was not conducted in English. There were a couple of prominent studies of subjects who spoke a language other than English that were cited many times in the literature. The difficulty of vowel patterns within different languages was a potential confounding variable. Definitionally,

because it was unclear how generalizable constructs for English were to other languages, difficulty was encountered when trying to code these studies. For example, the concepts of onset and rime become much more vague when dealing with an unfamiliar language. Studies omitted include "The Development of Analysis of Words into Their Sounds by Preschool Children" by Zhurova (1973) (Moscow), "Promoting Phonemic Analysis Ability Among Kindergarteners: Effects of Different Training Programs" by Cary and Verhaeghe (1994) (Lisbon, Portugal), "Evaluation of Long Term Effects of Phonemic Awareness Training in Kindergarten: Illustrations of Some Methodological Problems in Evaluation Research", by Olofsson and Lundberg (1985) (Umea, Sweden), and "Effects of a Training Program for Stimulating Skills in Word Analysis in First-Grade Children" by Lie (1991) (Norway).

Conclusion

The previous paragraphs described and defended the basis on which studies were eliminated from this meta-analysis. Studies were not omitted for a number of characteristics that varied across experiments. Studies were accepted regardless of whether training was explicit or implicit, whether groups were matched, and whether pretests and posttests were experimenter-designed or published materials. Studies were also accepted whether training was in simple (segmenting, blending, counting, isolating) or compound (deletion, substitution) phonemic awareness tasks (Yopp, 1988). The goal for the present study was to select original research that was similar enough to draw meaningful conclusions but varied enough to identify patterns in training procedures yielding the strongest effects.

Unless there was a distinct clue in the title or abstract of the study that an experiment was inappropriate for the meta-analysis, such as “dyslexics” or “remedial” or

"letter-sound training," all journal articles and ERIC documents that made the initial cut to 151 studies were located and reviewed before they were slotted for exclusion or inclusion. Because of access, decisions regarding dissertations were made based on abstracts and titles.

Eighteen studies made the final cut for inclusion. They were composed of 16 published experiments from professional journals, 1 dissertation, and 1 paper presented at a professional conference. Copies of all studies were obtained so they could be coded in preparation for analysis.

Coding Studies

Stock (1994) articulated the importance of the meta-analyst's understanding of the research base when developing a coding form. He explained, "A well-designed coding scheme is more likely if the synthesist knows both the research domain and research integration methods, because this knowledge provides the basis for making critical choices" (p.126). For this reason, coding form development did not begin until the literature on phonological awareness and on meta-analysis had been studied. In addition, coding form development was based on an actual selection of studies, as described below.

In order to prepare the coding form, a table of random numbers was used to select eight studies from those already obtained (Byrne et al., 1991; Content, Kolinsky, Morais, & Bertelson, 1986; McNeil & Coleman, 1967; O'Connor, Jenkins, Slocum, 1995; Olofsson & Lundberg, 1983; Rosner, 1971; Treiman & Baron, 1983; Weiner, 1994). The studies used to design the coding form were not necessarily used in the statistical analysis in this meta-analysis. These studies were read to determine which characteristics should be coded. In addition, an Associate Professor of Reading and a Professor of Educational

Research from the University of Kansas examined the coding form and contributed to its development.

Coding form development continued for a period of months. The coding form went through a number of revisions before it finally articulated the foci of this project. The evolution involved coding a few studies and then refining the form. This process was repeated approximately seven times. For one of the later revisions, several studies were coded with the dissertation advisor, a reading researcher. During all modifications of the coding form committee members assisted in clarifying definitions. All studies were coded with the final form, which was also used to calculate interrater reliability.

In reference to coding, Lipsey (1994) explained that "given the time-consuming and expensive nature of coding in research synthesis, the synthesist inevitably must find some balance between coding broadly for descriptive purposes and coding narrowly around the specific target issues of the particular synthesis" (p. 115-116). The goal for this meta-analysis was to achieve such a balance. The coding form has five types of information: general information, quality, subjects, training, and outcome measures. A copy of the coding form is in Appendix A. In order to clarify the connection between the coding form and the analysis, through the remainder of this report elements of the coding form that are later variables in the analysis are treated as proper nouns (i.e., they are capitalized).

Coding Form

Study identification. The first section of the coding form contains generic information which would be relevant to any meta-analysis. This section includes the title, author, article identification number, year of publication, source of data, and journal title.

Quality. The second section of the coding form gives a numerical value to the level of quality of the study. Each study can earn up to fifteen points based on measures taken against internal and external validity. This section addresses the data source, sampling procedures, sample size, assignment to groups, training, and testing. All studies were included in the statistical analysis regardless of their Quality Rating. Figure 1 is a graphic representation of the number of studies rated at a particular level of quality.

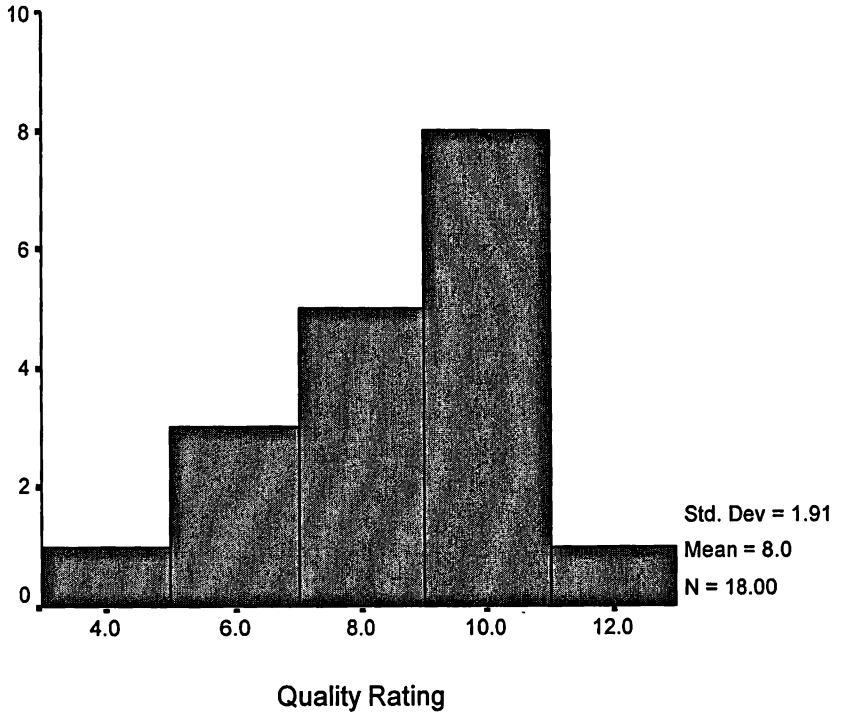


Figure 1. Quality of Studies

A second measure of internal and external validity was assigned by Troia (1999). Troia evaluated the methodological rigor of a body of phonemic awareness studies. He examined internal validity criteria such as, general design characteristics, measurement, and statistical treatment. His examinations of external validity included research hypotheses, participant selection and description, and generalization and maintenance measures. The evaluation resulted in a total score, Invalidity Index.

For studies included in this study that were examined by Troia, the Invalidity Index was also noted in the quality section of the coding form. The Quality Rating and the Invalidity Index shared some common gauges of validity, however Troia's Invalidity Index was more comprehensive. The correlation coefficient for Quality Rating and Invalidity Index was .08 and was not significant. The relationship is illustrated in Figure 2. As the Quality Rating increased, the Invalidity Index tended to increase. Only 12 data points are included in Figure 1 because Troia only assigned an Invalidity Index to 12 of the studies included in this meta-analysis.

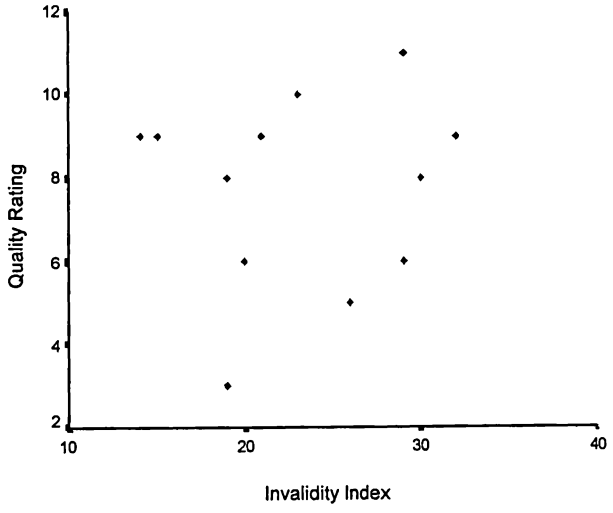


Figure 2. Correlation Between Quality Rating and Invalidity Index

Subjects. The third section of the coding form describes characteristics of the subjects. It states the total number of subjects in the study as well as numbers of subjects in each group. Within each group, this section also identifies the subjects' ages and Grade-levels at Training. The form can accommodate studies with up to three experimental groups and one control group.

Training. The fourth section is the one most specific to this particular meta-analysis. This section describes the training conducted in each group. First, the training section identifies the Approach of Training in phonemic awareness: segmentation, blending, addition, deletion, or other combinations. Because addition and deletion were only trained in conjunction with blending or segmenting, they were not examined separately in the statistical analysis but were labeled "other combinations". Within each of these approaches, the coding form details the particular task used in training. For example, the blending subgroup has tasks that involve the phoneme level and the onset/rime level.

The training section of the form also details whether students were taught to a certain performance criteria (Goal of Training) and whether letter/sound associations were incorporated into the training (Context of Training). Detailed definitions of the Training Approaches were used to train the second coder for reliability testing (see Appendix B). Finally, the training section of the coding form includes the number of students in the training group (Group Size) and the frequency and duration of training (Frequency of Training, Minutes per Training Session, and Total Minutes in Training). Using the coding form, training was described for all experimental and control groups for each study.

Outcome measures. Section V of the coding form presents the Outcome Measures which were used as posttests in the study. The Outcome Measures for phonemic awareness were divided into the following subgroups: segmenting, blending, deletion, and combination measures. See Appendix B for descriptions of Outcome Measures.

Interrater Reliability

Orwin (1994) stated that failure to gauge interrater reliability and to consider it when analyzing data can produce misleading results. For this study, interrater reliability was evaluated for each analyzed variable. This item-by-item calculation of interrater reliability is recommended by Orwin over across-the-board computations of reliability because, while the percentage of agreement may be high across all variables, it may vary considerably from variable to variable.

For all analyzed variables in this study, percent agreement (agreement rate) was calculated because it is the most commonly used index of interrater reliability (Orwin, 1994). Agreement rate is calculated by dividing the number of agreed upon observations by the total number of observations. For Quality Rating, the only variable with multiple data points, agreement rate was calculated by considering all the data points that contribute to the rating (9). This means that for the 10 studies checked for interrater reliability there were 90 data points. The raters were in agreement on 87 of those points (97% agreement).

The second rater for this study was a reading researcher who is familiar with both statistical methods and the phonemic awareness literature. Using form B, she was trained by the author and practiced on several studies until consistency between the two raters was established. Then, a random selection of 56% (10) of the studies was independently

rated by the second rater. Table 1 presents the variables rated and the corresponding percentages for interrater reliability. Interrater reliability for all variables ranged from 90-100%. Where coding differed between raters, the item was discussed until an agreement could be reached. The results of all coding of categorical and continuous variables are presented in Appendices C and D respectively.

Table 1

Interrater Reliability for Categorical and Continuous Variables

Categorical Variable	Agreement Rate	Continuous Variable	Agreement Rate
Approach of training	100%	Quality	97%
Goal of training	90%	Minutes per session	100%
Context of training	100%	Total minutes in training	90%
Outcome measures	100%	Frequency	100%
Grade level	100%	Group size	100%

Conclusion

The processes of searching the literature, defining the relationship to be studied, and coding studies each invite the introduction of bias to a meta-analysis. One defense, although insufficient on its own, is the inclusion of detailed descriptions of these critical steps. This section attempted to articulate the thought-processes and methods used in these preliminary steps which are so important to conducting a sound review. The following section describes the procedures used for conducting statistical analysis.

Statistical Analysis

In this study the independent variable was phonemic awareness training. The dependent variable was the phonemic awareness of the trained subjects. Moderator variables are those factors which may be capable of influencing the relationship between the independent and dependent variables. In this study, the moderator variables investigated were Approach of Training, Goal of Training, Context of Training, Grade-level at Training, Outcome Measures, Quality Rating, Invalidity Index, Minutes per Training Session, Total Minutes in Training, Frequency of Training, and Group Size.

Calculating Effect Sizes

For each study, an unbiased effect size, d^u, was calculated from the biased effect size, d. The effect size is the difference between the phonemic awareness training group and the alternate training or control group divided by the pooled standard deviation of the two groups (Cohen, 1977). The equation for d follows:

$$d = \frac{Ma-Mb}{SD}$$

where Ma and Mb represent the sample means of the comparison groups, and SD is the pooled standard deviation. Next, d is corrected for bias using the following equation:

$$d^u = c(m)d,$$

where c(m) is given approximately by the following equation:

$$c(m) \sim 1- \frac{3}{4m-1},$$

where m is the df computed from both the experimental and control groups (see Rosenthal 1994).

All effect sizes in this study were calculated from the means and standard deviations reported in the primary studies. Meta-analysis using the original means and

standard deviations gives a researcher an opportunity to seek out the most pertinent relationships within a study. For example, if a study had a phonemic awareness training group, an alternate training group, and a no-intervention control group, the relationship examined was that between the phonemic awareness and alternate training groups. In other words, whenever a control group with an intervention could be used over one without an intervention, the no-intervention control was not considered (B. -T. Johnson, personal communication, March 26, 1999). Each effect size was corrected to control for sample bias (Hedges & Olkin, 1985). A 95% confidence interval was drawn around each effect size. If the confidence interval (CI) does not include zero the researcher may conclude that there is a relationship between the dependent and independent variables.

Combining Effect Sizes

Given an awareness of the assumption of effect size independence, this study attempted to strike a balance between representing some studies heavily and investigating relevant questions. For the overarching research question (Does phonemic awareness training improve phonemic awareness?), analysis was first conducted at the study-level (with only one effect size per study). In each subsequent analysis, whether categorical or continuous, only one effect size from each of the relevant studies was used unless additional effect sizes contributed to the question under investigation. For example, if one study administered both blending and segmenting assessments, that particular study would contribute two effect sizes to the analysis of Outcome Measures.

For analysis of two variables, Approach of Training and Outcome Measures, 55% of the studies were represented more than once to allow for analysis of categories that varied within studies. For analysis of Grade-level at Training, one study was represented

twice. For all other categorical analyses and all continuous analyses, each study was represented once.

Effect sizes were combined to produce a weighted mean effect (d+) by weighting each effect size by the reciprocal of the within study variance. This procedure gives more weight to studies with larger samples and less variability among subjects. Next, a statistical test, Qw, was conducted to establish whether there was homogeneity of variance among effect sizes. The Qw statistic is approximately chi-square distributed with k - 1 degrees of freedom, where k is the number of studies (Hedges & Olkin, 1985). Qw indicates how well the combined effect sizes are represented by the weighted mean.

When the Qw term indicated that the effect sizes were not homogeneous, an analysis of outliers was conducted. Studies with extreme effect sizes, either positive or negative, were eliminated from the mean until an acceptable level of fit was achieved. After outliers were removed, the weighted mean was recalculated. The studies which produced outliers were examined to identify any methodological differences which may have caused the particular effect size to be extreme. Outliers were reinstated for model testing.

Model Testing

Categorical models. Categorical variables for this study included Approach of Training, Goal of Training, Context of Training, the Grade-level of Subjects at Training, and Outcome Measures. Categorical models were fitted to the effect sizes using Hedges and Olkin's (1985) methods. These procedures produce between-subgroup and within-subgroup effects. The Qb is a statistic which evaluates whether the mean effect size for different categories of a variable are the same. The between-subgroups effect is represented by Qb, which is approximately chi-square distributed with p - 1 degrees of

freedom, where p is the number of subgroups. The between-subgroups effect is analogous to a main effect in an ANOVA.

Within each subgroup, homogeneity was estimated by Qw, which is approximately chi-square distributed with m - 1, where m is the number of effect sizes in a subgroup. The Qw term evaluates whether the effect sizes for each category of a variable are homogeneous. For each subgroup of a categorical variable, a 95% confidence interval for the effect size was computed.

Continuous models. Continuous variables for this study were Quality, Invalidity Index, Minutes per Training Session, Total Minutes in Training, Frequency of Training, and Training Group Size. Continuous models are weighted least squares regressions examining one predictor at a time. They are calculated by weighting each effect size by the reciprocal of its variance. Estimation of each regression model yields a statistical significance test of each predictor and a test of overall model specification evaluating whether significant systematic variation remains unexplained (Hedges & Olkin, 1985). The test of model specification utilizes the sum of squares error statistic, Qe, which is approximately chi-square distributed with k - 2 degrees of freedom where k is the number of effect sizes (excluding the intercept). If Qe is significant it implies heterogeneity of effect size after taking into account the predictor.

Conclusions

This chapter has described the methodology used in the current study. It began by presenting the research questions. Next, it detailed the procedures for locating studies, identifying studies for inclusion, and coding studies for analysis. Finally, it described the procedures that were used in the statistical analysis. This included explanations of the calculation and interpretation of individual and combined effect sizes. Chapter III also

described the procedure for analyzing moderator variables both categorically and continuously.

CHAPTER IV

Results

The following analysis investigates the relationship between the independent variable, Phonemic Awareness Training, and the dependent variable, Phonemic Awareness. Further analyses probe how moderator variables affect this relationship. Presentations of the results deal with 1) the overarching research question exploring the relationship between training and the dependent variable, 2) categorical moderator variables, and 3) continuous moderator variables.

Statistical Analysis

Overall Effect Sizes

In the current meta-analysis, Phonemic Awareness Training was found to significantly affect Phonemic Awareness. By Cohen's definition (1988), the overall effect size for Phonemic Awareness was strong ($\underline{d+}$=1.23). The effect sizes ranged from .32 to 3.46 with a median of 1.50.

Outlier diagnosis. Outlier diagnosis was conducted to determine the homogeneity among studies. Because the $\underline{Qw}$ statistic gauges heterogeneity, its significance indicates a lack of homogeneity. Five outliers had to be removed for homogeneity to be achieved. The $\underline{Qw}$ term before and after outlier diagnosis was 76.28, $p<.001$ and 20.27, $p=.06$, respectively. For Phonemic Awareness, the overall mean after outliers were removed ($\underline{d+}$=1.57) was higher than that before. This indicated that the majority of outliers had low rather than high effect sizes. This is demonstrated in Table 2 which is a chart of the effect sizes and outlier diagnosis. Figure 3 presents a visual display of effect sizes before and after outlier diagnosis, graphically presenting the tendency for low outliers.

Table 2

Study-level Phonemic Awareness Effect Sizes and Outlier Diagnosis

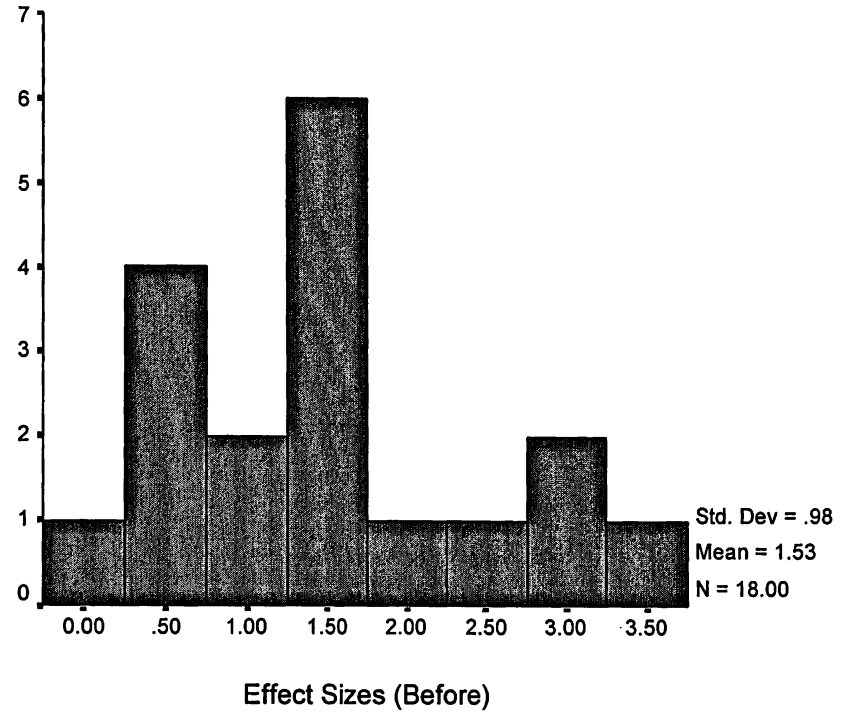
Study #	Training n	Control n	d^u	95% CI
2	12*	12	2.74	1.63 to 3.86
4	5*	12	0.34	-0.71 to 1.39
5	3*	3	2.95	0.64 to 5.26
6	5*	5	3.46	1.50 to 5.41
8	8	8	1.51	0.40 to 2.62
9	18	17	1.43	0.69 to 2.18
10	8*	4	1.74	0.36 to 3.13
13	20	11	1.14	0.35 to 1.93
14	56	28	1.67	1.15 to 2.19
15	84	75	1.77	1.40 to 2.13
18	29	30	1.49	0.91 to 2.06
19	40	20	1.07	0.50 to 1.64
20	45	23	1.69	1.11 to 2.26

Table 2 (continued)

Study #	Training n	Control n	d^u	95% CI
		Outliers in Order of Removal		
16	58	25	0.19	-0.28 to 0.66
7	30	10	2.91	1.95 to 3.87
17	60	40	0.71	0.30 to 1.12
11	18	18	0.32	-0.34 to 0.97
12	21	21	0.42	-0.20 to 1.03

*n represents groups used in analysis rather than sample size.

Note. Studies corresponding to study numbers are listed in Appendix E.



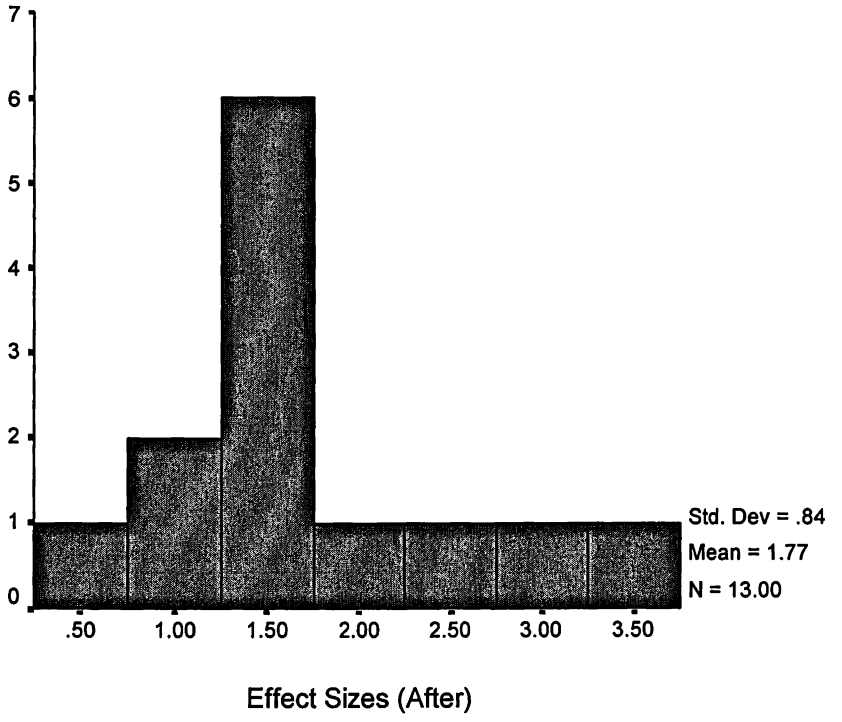


Figure 3. Effect Sizes Before and After Outlier Diagnosis

Relationship Between Effect Size and Sample Size. The relationship between the effect sizes and the sample sizes was investigated. The correlation between the two was negative but not significant (r=-.19, p=.457). A scatterplot (Figure 4) confirmed that effect sizes tended to increase as sample sizes decreased.

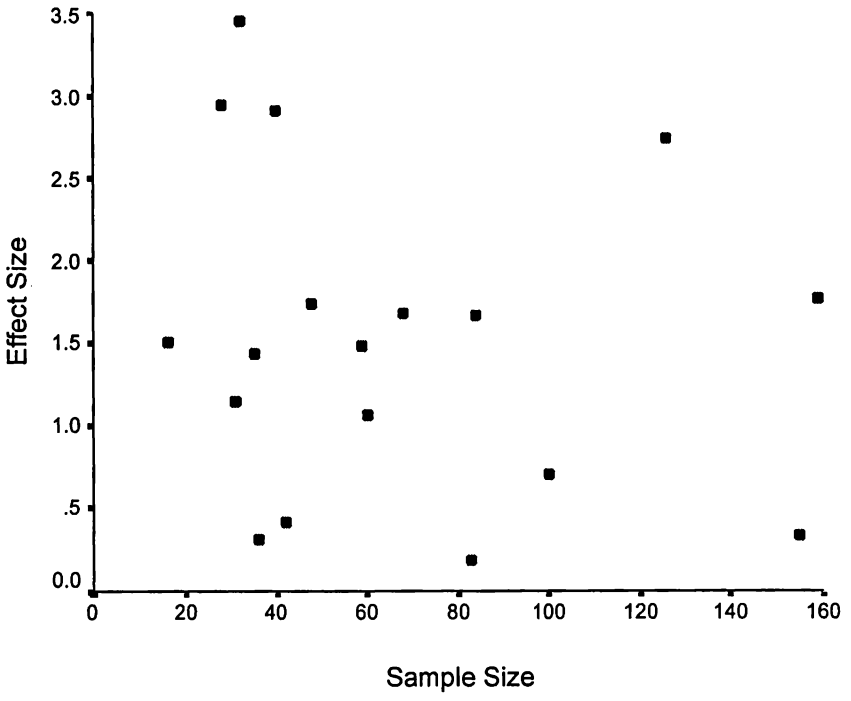


Figure 4. Effect Sizes and Sample Sizes

Categorical Variables

Categorical model testing was conducted on the following variables: Approach of Training, Goal of Training, Context of Training, Grade-level at Training, and Outcome Measures. The results of all categorical analyses are presented in Table 3. Table 3 also presents the statistical term Qb, which determines significance between subgroups. As presented in Table 3, within variables, the difference between subgroups was significant for Approach of Training and Outcome Measures.

Table 3

Statistical Analysis for Categorical Variables

Variable	Qb	p
Approach	15.45	.001
Goal	0.36	.550
Context	0.02	.887
Grade	3.40	.183
Outcome measures	18.03	<.001

Table 4 presents the effect sizes, confidence intervals, and homogeneity terms (Qw) for subgroups of variables. Because Approach of Training and Outcome Measures were the two variables which indicated a significant difference between subgroups, subgroup examination is warranted. The strongest mean effect size for Approach of Training was associated with segmenting/blending training (d==1.41). The weakest effect size for Approach of Training was for combination measures (d+=.95). For Outcome Measures, the strongest and the weakest effect sizes were

associated with combination measures and deletion respectively. No other categorical variables had significant differences between groups.

Mean effect sizes (d+) for all subgroups were significantly different from zero, as indicated by the associated confidence intervals. The only Qw term that indicated homogeneity was associated with the subgroup first-grade within the variable of Grade-level at Training. However, the homogeneity may be related to the fact that there were only three cases in that subgroup.

Table 4

Effect Sizes, Confidence Intervals, and Homogeneity Terms for Subgroups of Variables

Subgroup	# of d^u	d+	95% CI	Qw	p
			Approach		
Segmenting	15	0.97	0.79 to 1.14	118.73	<.001
Blending	6	1.07	0.61 to 1.53	44.30	<.001
Seg/blend	17	1.41	1.23 to 1.59	58.52	<.001
Combinations	11	0.95	0.73 to 1.18	50.39	<.001
			Goal		
Mastery	7	1.04	0.77 to 1.31	24.00	.001
Non-mastery	11	1.33	1.14 to 1.51	49.30	<.001
			Context		
Letters	7	1.24	0.04 to 1.45	36.19	<.001
No letters	11	1.22	0.99 to 1.45	40.07	<.001
			Grade-level		
Pre-K	3	1.45	0.92 to 1.98	9.47	<.05
Kindergarten	13	1.26	1.09 to 1.43	60.21	<.001
First-grade	3	0.87	0.43 to 1.31	5.02	.170

Table 4 (continued)

Subgroup	# of d^u	d+	95% CI	Qw	p
		Outcome Measures			
Segmenting	13	1.04	0.90 to 1.18	154.92	<.001
Blending	11	1.40	1.16 to 1.65	72.84	<.001
Deletion	5	0.74	0.42 to 1.06	16.65	<.05
Combination	6	1.52	1.21 to 1.82	24.96	<.001

Follow-up analyses. Because two variables (Approach of Training and Outcome Measures) containing four subgroups were significant between subgroups, pairwise comparisons were conducted to identify which relationships within these variables were significant. Significance levels were set using the Holm (1979) procedure. Table 5 presents these results. For Approach of Training, only two comparisons were significant. For Outcome Measures, four comparisons were significant.

Table 5

Pairwise Comparisons for Phonemic Awareness Variables

Comparison	chi square	p
	Approach	
Segmenting vs. seg/blend	12.02	.008
Seg/blend vs. other combinations	10.08	.01
Blending vs. seg/blend	1.82	ns
Blend vs. other combinations	0.21	ns
Segmenting vs. blending	0.16	ns
Segmenting vs. other combinations	0.02	ns
	Outcome Measures	
Deletion vs. combination	11.72	.008
Blending vs. deletion	10.48	.01
Segmenting vs. combination	7.46	.012
Segmenting vs. blending	6.30	.016
Segmenting vs. deletion	2.93	ns
Blending vs. combination	0.30	ns

Note. For all significant comparisons, the subgroup with the strongest effect is underlined.

Variation among effect sizes. The proportion of variation among effect sizes that is accounted for by each variable is represented by r square. See Table 6. The most variation among effect sizes was accounted for by Outcome Measures. The least variation among effect sizes was accounted for by Goal of Training (letters vs. no letters) and Context of Training (letters vs. no letters). Each of the variables accounted for less than 10% of the variance among effect sizes.

Table 6

R Square for Categorical Variables

Variable	r square
Approach	.05
Goal	.00
Context	.00
Grade	.04
Outcome measures	.06

Continuous Variables

Weighted least squares regressions (Hedges & Olkin, 1985) were conducted to determine whether moderator variables influenced the relationship between training and phonemic awareness. The continuous moderator variables explored were Quality, Invalidity Index, Minutes per Training Session, Total Minutes in Training, Frequency of Training, and Size of Training Group.

Table 7 presents the beta weights, z-values, and homogeneity terms (Qe) for continuous variables. Except for Quality Rating, all continuous variables were negatively associated with effect sizes, as indicated by their beta weights. Quality Rating was positively and significantly related to effect sizes, as indicated by the overall regression effects and their associated p values. Figures 5-9 illustrate the significant relationships between moderator variables and phonemic awareness effect sizes. For all regressions, Qe indicated that unexplained variation remaining in the model after controlling for a given predictor was significant; the differences in the effect sizes were not adequately explained by the predictors.

Table 7

Beta Weights, Overall Regression Effects, and Homogeneity Terms for Continuous Variable Analysis

Predictor	# of d	Beta Weight	z-value	p	Qe	p
Quality	18	.13	2.71	.006	70.13	<.001
Invalidity	13	-.06	3.10	.002	36.76	<.05
Min./session	18	-.07	3.49	<.001	65.47	<.001
Total minutes	19	-.00	4.09	<.001	60.73	<.001
Frequency	18	-.07	1.05	.292	76.37	<.001
Group size	18	-.05	4.63	<.001	56.16	<.001

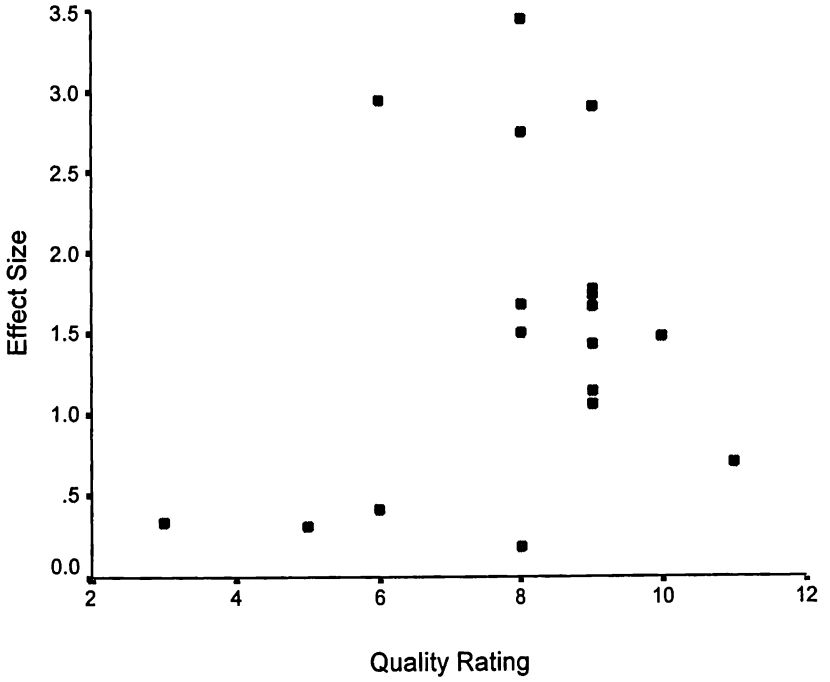


Figure 5. Effect Sizes and Quality Rating

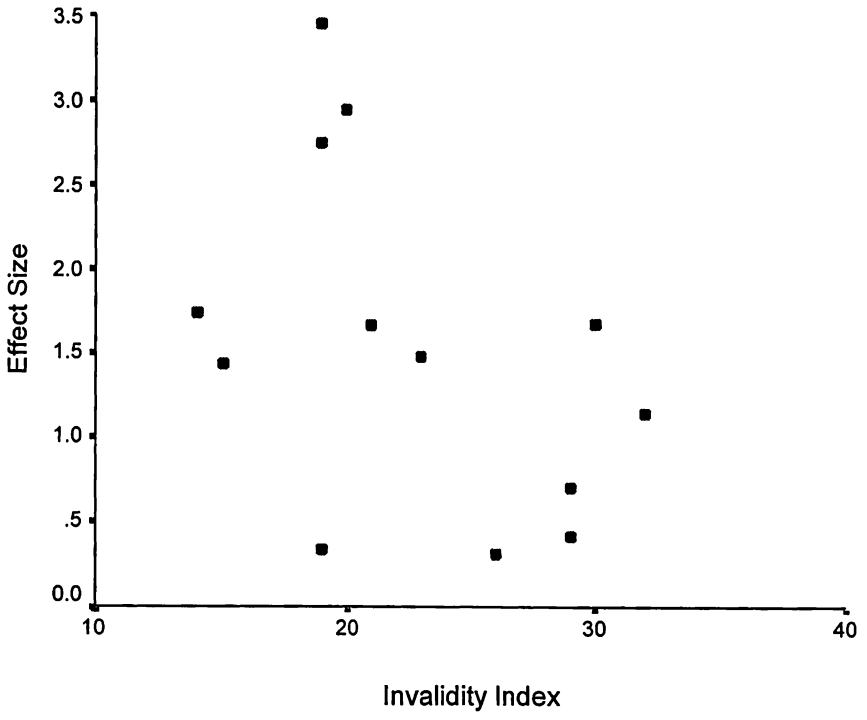


Figure 6. Effect Sizes and Invalidity Indices

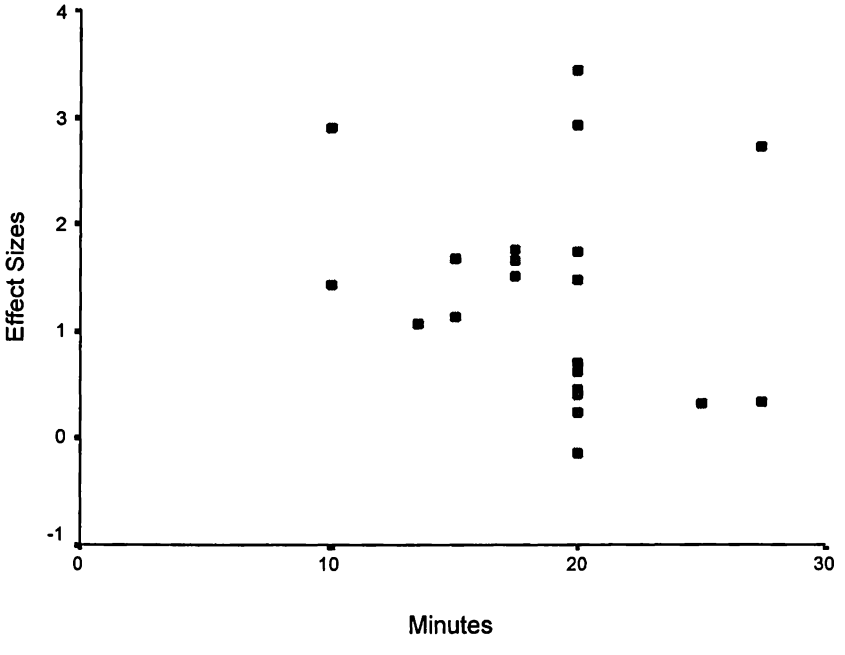


Figure 7. Effect Sizes and Minutes per Training Session

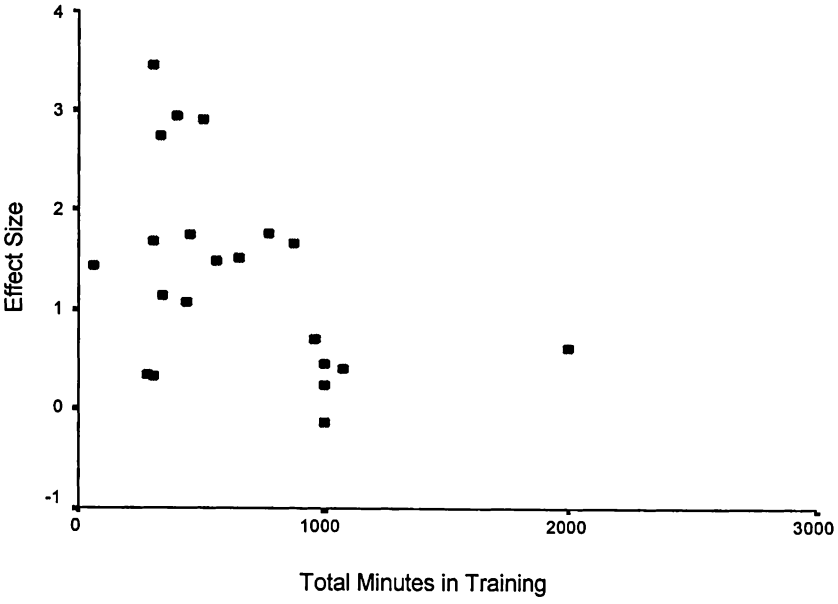


Figure 8. Effect Sizes and Total Minutes in Training

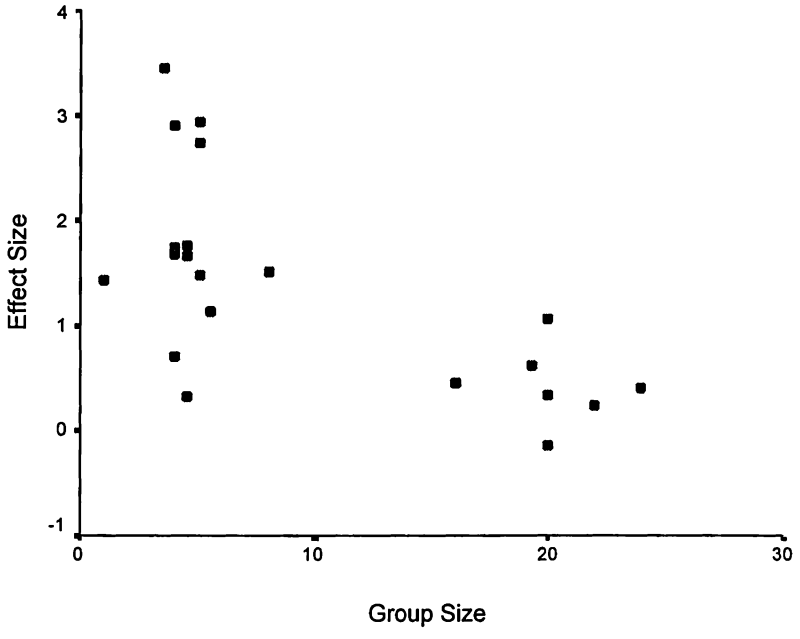


Figure 9. Effect Sizes and Training Group Sizes

Variation among effect sizes. To determine what proportion of the variance among studies is accounted for by particular variables, r square were calculated for each continuous variable. See Table 8. The highest r square was associated with Group Size, indicating that 28% of the variability among effect sizes is accounted for by the size of the training group. Frequency of Training had the lowest r square value. The r square for continuous variables tended to be stronger than those for categorical variables.

Table 8

R Square for Continuous Variables

Variable	r square
Quality	.09
Invalidity	.20
Minutes/session	.10
Total minutes	.22
Frequency	.01
Group size	.28

Follow-up analyses. As previously reported, effect sizes were negatively and significantly correlated with the continuous variable of Minutes per Training Session and Total Minutes in Training. As training time increased effect sizes tended to decrease.

In investigating these unusual results, correlations between Group Size and Minutes per Training Session and between Group Size and Total Minutes in Training were conducted. The correlation between Minutes per Training Session and Group Size was not significant (r=.23, p>.05). Thus, Group size was not examined as contributing to the peculiar findings for Minutes per Training Session. The correlation between Total Minutes in Training and Group Size was positive and significant (r=.50, p<.05) .

Because Total Minutes in Training was significantly correlated with Group Size, a weighted least squares regressions was conducted to determine if results associated with time were a product of size of training groups rather than time in training. The regression demonstrated that both Total Minutes in Training (z=2.65, p=.007) or Group Size (z=3.41, p=.001) remained significantly related to effect size when the other was taken into account.

Conclusions

The overall mean effect for this study indicated that there is a strong relationship between phonemic awareness training and phonemic awareness. The removal of five outliers from among the effect sizes was required to achieve homogeneity.

After reinstating outliers, moderator variables were analyzed. The effect sizes for subgroups within Approach of Training were found to be significantly different from each other. The differences in the subgroups in Goal of Training, Context of

Training, and Grade-level at Training were not significantly different from each other. Outcome Measures provided significant differences favoring combination measures. This was confirmed with post hoc analyses.

Weighted least squares regressions were conducted with continuous moderator variables. Effect sizes were significantly and negatively correlated with the Invalidity Index, Minutes per Training Session, Total Minutes in Training, and Group Size. The relationship between Quality Rating and effect size was positive and significant.

Group Size and Total Minutes in Training were significantly and positively correlated. A follow-up weighted least squares regression demonstrated that Total Minutes in Training and Group Size remained significant after taking into account the other.

CHAPTER V

Discussion

This chapter begins with a discussion of results related to the overarching research questions: Does training in phonemic awareness impact phonemic awareness? Next, insights into the effects of moderator variables on effect sizes are presented. After the results discussion, Chapter V details a number of threats to the validity of this meta-analysis. Each point is weighted against the strategies enlisted to protect this study from that particular threat. The chapter ends with implications this research has for the classroom and suggestions for future research in phonemic awareness training.

Overall Effects

Just from exposure to the original literature, few reading researchers would deny that phonemic awareness can be trained. In fact, the relationship between phonemic awareness training and phonemic awareness acquisition has generally been accepted by reading researchers (Adams, 1990; Snow, Burns, & Griffith, 1998). However, the statistical aggregation of the Phonemic Awareness Training effect sizes is helpful for solidifying researchers' understanding of the relationship between phonemic awareness training and phonemic awareness.

If a large effect size by Cohen's (1988) definitions is .80, the overall effect size for this study, 1.23, is clearly strong. If not for the heterogeneity repeatedly found within subgroups, this solid result would make a virtually inarguable case that phonemic awareness can be increased with training. The heterogeneity implies that, under some conditions, the effect size may be some degree lower. However, the effect size is high enough to withstand this concern.

Miller and Pollock (1994) define an empirical law as a "confirmed hypothesis--a relationship between two concepts that has been (relatively) well established by research" (p.458). By this definition, the relationship between training in phonemic awareness and phonemic awareness acquisition approaches empirical law.

Outlier diagnosis

In examining the studies removed during outlier analysis, some commonalties were identified. Using the coding summary tables in Appendices D and E, the outliers were studied. Two issues arose. First, of the five outliers removed, three were the studies that spent the most total time training students (Ayres, 1996; Brady, Fowler, Stone, & Winbury, 1994; Torgesen, Morgan, & Davis, 1993). This theory does not hold true for the other two studies, however, as one falls at the median on total training time (Davidson & Jenkins, 1994) and one falls below the median (Weiner, 1994).

A second interesting point in regard to the outliers is that three of the five (Brady et al., 1994; Davis et al., 1993; Weiner, 1994) have the highest Invalidity Indices. The remaining studies were not rated. However, it is possible that methodological flaws separate these studies from the body of homogenous studies that remained after these were removed.

The source of the variability within the studies cannot be completely determined. However this moderate amount of heterogeneity is not enough to disregard the findings. Eagly and Wood (1994) make this point: "In general, findings that are homogeneous except for a few outlying effect sizes also suggest that further research in the same paradigm would be redundant" (p.487).

In addition, the homogeneity statistic is not just a function of variation between studies, it is also a function of sample size. Increasing the sample size increases power, thus increasing the likelihood that a statistical test will be significant. Consequently, the large sample size associated with this study may have affected the homogeneity statistic.

Moderating Variables

Categorical Variables

Approach of Training. The difference in subgroups within Approach of Training was significant. Segmenting/blending was found to have the strongest effect. This result is particularly valid because some of the studies in this collection examined this relationship within the study. Consequently, this result is a product of between study and within study examinations. The studies included in this review which compared segmenting and blending training to blending only and/or segmenting only are described in the following paragraphs.

Davidson and Jenkins (1994) utilized three experimental groups (segmenting training only, blending training only, segmenting and blending training) in their study of phonemic awareness training. Subjects were subsequently assessed on measures of segmenting and blending. The researchers found that subjects trained to segment and blend were able to perform significantly better on measures of segmenting and on measures of blending than subjects only trained in one aspect of phonemic awareness. Fox and Routh (1984) compared segmenting training to segmenting/blending training. Their results confirmed those of Davidson and Jenkins; Fox and Routh found that subjects trained in segmenting/blending performed significantly better on a segmenting measure and on a blending measure. Torgesen and his associates (1992) compared a blending training only with segmenting and blending training. Their

results are supportive of those previously described; subjects trained in segmenting and blending did significantly better than the language-experience control group while those trained only in blending did not.

In the current study, follow-up analyses demonstrated that segmenting/blending training was significantly better than segmenting training or other combinations training. However, segmenting/blending training was not significantly better than blending training. While segmenting/blending training produced significantly higher effects than the other trainings, it is interesting to notice the magnitude of effect sizes associated with the other subgroups because the phonemic awareness effect sizes for all Approaches of Training are high by Cohen's (1988) definition and significant. In fact, the weakest mean Phonemic Awareness effect size for Approach of Training is .95.

While there is within and between study confirmation of the advantage of training segmentation and blending jointly, there is conflicting evidence within studies as to whether this advantage holds true for reading and spelling achievement. For example, Davidson and Jenkins (1994) found an advantage in word learning on an analogue test for students trained only in segmenting over students trained in segmenting/blending. They did not find a reading advantage for students trained only in blending. Fox and Routh (1984) found opposite results with their reading analogue tasks; they found training in segmenting alone to be insufficient to "enable the child to decode written words to speech" (p.1063). Finally, Torgesen and his colleagues (1992) found that students trained in segmenting and blending made significantly fewer errors on an analogue task than those trained in blending only.

While the within-study results for Approach of Training are congruent in regards to phonemic awareness, they are less consistent for reading achievement.

Given that phonemic awareness is of limited value in itself but rather is important as a correlate with reading ability (Ehri and Wilce, 1985; Fox & Routh, 1976; Juel, 1988; Tunmer & Nesdale, 1985), further investigation of this conflict is merited.

Goal of Training. There was not a significant difference between training that focused on mastery and in training that focused on covering the lesson. Some may consider this result unexpected. One might assume that students taught to mastery would outperform students who were not. However, the research on a related topic, mastery learning, has produced mixed results.

Slavin (1987) conducted a "best-evidence synthesis" of the research associated with mastery learning in elementary schools, only including studies that trained students for a minimum of 4 weeks. Slavin found that there were no effects for mastery learning on standardized assessments and there were only moderate effects for experimenter-designed measures. Kulik, Kulik, and Banger-Drowns (1990) arrived at an opposite conclusion when they conducted a meta-analysis of 108 mastery learning effect sizes. They found that mastery learning had positive effects on the performance of children in upper elementary school. Slavin (1990) aptly defended his position and argued, "The claim that mastery learning can accelerate achievement in elementary and secondary schools is still awaiting convincing evidence" (p. 301).

One argument Slavin (1990) posits is that positive effects demonstrated by mastery learning may simply be a product of learning shifts; low achievers learn while typically high achievers may not perform as well. This "Robin Hood" effect, taking from the high achievers and giving to the low achievers, is given some credence because in several of the studies Slavin reviewed, effects were only high for low achievers. If this phenomenon is true, it would be muted in the present meta-analysis

because most of the subjects were of average ability. This may explain why the results of this study do not manifest the perceived benefits of mastery teaching.

There is a second plausible explanation for the absence of a significant difference between students taught with a mastery emphasis and students taught with a lesson emphasis. Because teaching to mastery criterion generally means teaching longer, the results for Goal of Training are related to the significant time variables presented in the continuous analysis (Minutes per Training Session & Total Minutes in Training). With most of the studies in this analysis, teaching to mastery involved giving extra lessons to those students who needed them. However, this meta-analysis indicates that more time in training may be less effective; time seems to be associated with something that interferes with learning. The unusual relationship between time in training and effect size is thoroughly explored in the discussion of the continuous variables related to time.

Context of Training. Based on the results of this meta-analysis, whether sound/letter relationships are taught in conjunction with phonemic awareness training does not affect phonemic awareness acquisition. However, this meta-analysis does not explore the effects of phonemic awareness training on learning to read. While adjunct sound/letter training may not extend phonemic awareness, it may produce benefits in reading or spelling achievement.

Addressing the Context of Training (sound/letter instruction vs. none) leads to the circular question which has plagued phonemic awareness researchers: Does phonemic awareness lead to reading success or does beginning reading skill lead to awareness of phonemes? This question is complex and conflicting research on the subject has led to strong debate.

Knowledge of letter names is highly correlated with success in beginning to read. In fact, in their classic study of first-grade classrooms across the United States, Bond and Dykstra (1967) found that knowledge of letter names before formal reading instruction was the strongest predictor of the reading success of first-grade students. Since then, the correlation between letter names and beginning reading success has been well-documented (Speer & Lamb, 1976; Stanovich, Cunningham, & Cramer, 1984; Tunmer, Herriman, & Nesdale, 1988). The second strongest predictor of success in first grade found by Bond and Dykstra was phonemic awareness. This correlation has also been thoroughly established (Juel et. al, 1986; Lomax & McGee, 1987; Swank & Catts, 1991). Consequently, one might expect a positive correlation between phonemic awareness and knowledge of letter names. However, this is not consistently true.

Naslund and Schneider (1996) specifically addressed the question of whether knowledge of letters was necessary for development of phonological awareness. This longitudinal study demonstrated that

> The hypothesis that children cannot develop phonological awareness before they are familiar with individual grapheme-phoneme correspondences...does not appear to be supported here. The majority of children in this study with high phonological awareness at the same time demonstrated low letter knowledge. (p. 55)

Furthermore, those tasks which predicted reading significantly did not require specific knowledge of sound/letter correspondences. Promoting the same argument, Tunmer and Rohl (1991) maintain that in order for children to benefit from early instruction in phoneme/grapheme relationships, they must first possess a basic consciousness of the sounds within words.

Conflictingly, Byrne and Fielding-Barnsley (1991) found that students who demonstrated skill with letter knowledge and phoneme identity were more successful at word reading than students who demonstrated competence in either skill alone. In addition, it is interesting that adults who are only proficient readers of Chinese characters, with no experience with an alphabetic script, are unable to delete or add phonemes to spoken words. In contrast, comparable readers of both Chinese characters and Chinese alphabetic spellings were adept at performing these phonemic manipulation tasks (Read, Zhang, Nie, Ding, 1986).

It appears that this chicken-and-egg debate will persist. Perfetti (1984) logically argues that some awareness of phonemes enhances initial reading growth and vice versa; each skill builds the other and produces reciprocal growth. Nausland and Schneider (1996) suggest that phonological processes vary; some develop spontaneously without knowledge of letters while others require some facility with letters and their corresponding sounds. Stahl and Murray (1994) further this argument, suggesting that alphabet knowledge precipitates the ability to segment onsets from rimes, and that this particular segmenting skill is a prerequisite for word reading and complex phonemic awareness.

Additional research, both primary studies and literature syntheses, are needed to establish and define the connections between phonemic awareness training with adjunct sound/letter instruction and phonemic awareness, reading achievement, and spelling achievement. The larger issue of whether it is necessary for one skill (phonemic awareness or knowledge of sound/letter correspondences) to precipitate the other also invites further investigation.

Finally, for the current meta-analysis, the lack of connection between sound/letter incorporation and improved phonemic awareness (and possibly reading

and spelling achievement) may be due to the limitations of studies included in this synthesis. When sound/letter instruction supplemented phonemic awareness training in the studies included for review, it did so minimally. In fact, if a training emphasized phoneme/grapheme relationships it was excluded from this study. The result of the analysis of Context of Training may simply indicate that sound/letter instruction, quite appropriately in terms of this synthesis, did not receive much time or attention during training.

Grade-level at Training. There were no significant differences between subgroups for Grade-Level at Training. The absence of significance may be due to the lack of studies in the prekindergarten and first-grade subgroups. These two subgroups had only three studies. Although not significant, the differences between subgroups remains interesting because the confidence intervals indicated that the effect sizes for each subgroup were significant. Interestingly, the largest effect size produced from Grade-level at Training analyses was associated with prekindergarten subjects. While only representing three studies (Byrne et al., 1991; Byrne et al., 1995, Experiment #2, Slocum, O'Conner, & Jenkins, 1993), the aggregate effect associated with prekindergarten is worth considering because it was significant and because prekindergarten seems to be overlooked in the phonemic awareness training literature.

The first prekindergarten study considered in this meta-analysis was conducted by Byrne and Fielding-Barnesley (1991) and produced a very strong effect size ($\underline{d}$=2.68). In this study 64 preschool subjects were trained in "Sound Foundations" (Byrne & Fielding-Barnesley, 1991). Using a variety of posters, games, and worksheets, this program teaches students to discriminate initial and final phonemes. After 12 weeks of training, students were assessed on initial and final

phoneme identity. They were asked to identify, from three pictures, which picture started or ended the same as a word pronounced by the examiner.

Byrne and Fielding-Barnesley conducted follow-up analyses with the original subjects when they were in kindergarten (1993), first-grade (1995), and second-grade (1995). This valuable data on preschool phonemic awareness training is not captured in the current research synthesis as the longitudinal reports examined reading and spelling achievement rather than phonemic awareness. Byrne & Fielding-Barnsley found that children who had developed phonemic awareness demonstrated advanced literacy development during first grade. Also, this advantage remained when letter knowledge was statistically considered. Throughout the longitudinal data, the authors found superior decoding skills in students who were trained in phonemic awareness. There was also evidence that the experimental subjects were ahead of the control subjects on measures of reading comprehension.

The second prekindergarten study included in this analysis was also conducted by Byrne and Fielding-Barnsley (1995, Experiment #2). This study also trained prekindergarten children in the "Sound Foundations" program. The resultant effect size for the study was moderate ($\underline{d}$=.36). The difference between the longitudinal study effect size and the effect size associated with the later study was possibly due to the very different training group sizes. Subjects were trained in groups of five in the first study while they were trained in intact classrooms in the second study.

Finally, Slocum and his associates (1993) trained Headstart children to segment and blend onsets and rimes. This study was divided into two phases which examined the transfer of skills from segmenting to blending and vice versa, as well as the effects of training on performance. In terms of transfer of phonological processes from one to the other, there was no indication that segmenting ability was transferable

to blending ability or vice versa. In terms of training effects, the study produced a very strong effect (d=1.47), which represented an average of segmenting and blending assessments. These results are intriguing considering the limited number of prekindergarten studies available for examination and the socioeconomic status of the subjects. Slocum et al. stated, "These large effects are particularly notable considering the age and background skills of the children who participated in the study" (p. 626).

In summary, all of the subgroups within Grade-level at Training had significant average effect sizes; however, these effect sizes were not significantly different from each other. Because there are a limited number of preschool effect sizes available for synthesis, it is difficult to draw conclusions about the most effective grade for phonemic awareness instruction. However, the effects associated with prekindergarten are intriguing because they are higher than one might expect, even with relatively brief training periods, and because phonemic awareness training research tends to focus on kindergarten subjects. Additional phonemic awareness training research conducted with prekindergarten subjects is needed.

Outcome Measures. The difference in the effect sizes of subgroups within Outcome Measures was significant, and each subgroup mean achieved significance individually. Combination measures demonstrated the strongest effect. The second strongest effect was associated with blending tasks and the weakest effect was associated with deletion tasks.

The fact that combination measures had the strongest effect is interesting, considering that combination measures includes substitution tasks. Substitution tasks are considered particularly difficult (Yopp, 1988). Adams (1990) wrote of such assessments:

> Tasks of this ilk have generally been found to be beyond the reach of children before the very end of first grade. This should not be surprising. . . . To pick out the relevant phoneme(s) from any given test word, the children must have well-developed phonemic segmentation skills. Then, whether to delete or reorder the phonemes or insert a new one and put the new word back together, must require all manner of memory skills and gymnastics. It is hard to imagine how one might succeed in such tasks without fairly well developed spelling skills. (p. 72)

Given these results, it seems relevant to examine the individual studies which contributed to the weighted mean for combination measures.

The mean for combination measures represented an aggregation of six effect sizes from only four studies (Castle, Riach, & Nicholson, 1994, Experiment #1; Cunningham, 1990; O'Conner et al., 1995; Weiner, 1994). Two of the studies (Castle et al. & Weiner) used substitution and deletion tests. The other two studies used the Lindamood Auditory Conceptualization Test (LAC) (1979) which requires children to use colored blocks to represent the number and order of perceived phonemes in pronounced words.

Predictably for kindergarten, the effect size for the study conducted by Weiner (1994) was quite low ($\underline{d}$=.18). This effect size represented an average of effects produced by tests of deletion, deletion and substitution of initial phonemes, and deletion and substitution of final phonemes. The training group did not perform significantly better than the control group on any of the measures.

The Castle et al. study (1994) produced an abnormally large effect size ($\underline{d}$=3.68), particularly considering the measure and the age of the subjects. Castle administered Roper's (1994) phonemic awareness measure to children in the first few

months of kindergarten. This measure consists of subtests including segmenting, blending, deletion of initial phonemes, deletion of final phonemes, substitution of initial phonemes, and substitution of final phonemes. These findings contradict what previous research has indicated about such complex training tasks (Rosner, 1974; Yopp, 1988)

The other two studies aggregated in Combination Measures (Cunningham, 1990; O'Conner et al., 1995) used the Lindamood Auditory Conceptualization Test (1979). They contributed four effect sizes, the lowest of which was 1.61. The strength of these effects may be due to the training. In both studies training involved using manipulatives to represent the sounds in words. Such training parallels the Lindamood Auditory Conceptualization Test in that the latter requires students to manipulate colored blocks to represent the number and order of phonemes in a word. The high effects associated with these studies may be connected to the close alignment between the training and the assessment.

Given the unusual effects associated with combination measures, the second largest effect size ($\underline{d}+=1.40$), also merits investigation. Follow-up analyses indicated that the two strongest subgroups for Outcome Measures (combination measures and blending) were not significantly different from each other. The mean effect associated with blending measures was an aggregation of 11 effect sizes from 6 different studies. The relevance of blending measures is enhanced by the support of other research. Perfetti, Beck, Bell and Hughes (1987) found that children develop competence in phoneme blending before they develop competence in phoneme deletion. In her analysis of phonemic awareness assessments, Yopp (1988) arrived at a similar conclusion. She stated, "Phoneme blending is one of the easier phonemic awareness tests for young children" (p. 171).

In sum, while combination measures produced the largest weighted mean effect size for Outcome Measures, the limited number of studies in this subgroup and the extreme score in favor of substitution tasks in early kindergarten cast a shadow of doubt on the result. On the other hand, the blending Outcome Measures, which was not significantly different than combination measures, appears to be more reliable as it is supported by other research.

Continuous Variables

Quality and Invalidity Index. The second section of the coding form assigned a Quality Rating to each study. See Appendix A. The relationship between Quality Rating and effect size was positive and significant. Also of interest, the Invalidity Index (Troia, 1999) was negatively and significantly related to phonemic awareness effect sizes. As the Invalidity Index decreased, the effect sizes increased significantly.

Considering the r square for Invalidity Index (.20) and Quality Rating (.09), one can conclude that Invalidity Index is a slightly stronger predictor of effect size. Examinations of the two indices revealed that everything included in the Quality Rating was also included in the Invalidity Index. However, the Invalidity Index is much more comprehensive, including 26 additional criteria for gauging validity. One could conclude that some of these additional criteria may be causing the Invalidity Index to be a stronger predictor of effect size. However, determining to which variables the difference is attributed would require regressions between the effect sizes and each of the items gauged by the Invalidity Index. These analyses are not a part of this study.

Variables associated with time. Total Minutes in Training and Minutes per Training Session were negatively and significantly related to effect sizes. In addition, Total Minutes in Training remained significant after controlling for Group Size;

Minutes Per Training Session was not correlated with Group Size. These results were unusual because they indicated that more training does not necessarily lead to higher proficiency in discriminating and manipulating phonemes. There seems to be a point at which the effects level out or decline with longer or additional training sessions.

There are a number of plausible theories that could explain this occurrence. First, perhaps researchers who train more frequently or for longer durations are also doing something that inhibits learning. For example, trainers who train for long durations may venture from a specific lesson format as they become comfortable with it. Another possibility is that trainers who are expected to accomplish an objective in a few weeks may be more efficient and have higher expectations of students than trainers who are expected to accomplish a similar task over a period of months.

A second possible explanation for the inverse relationship between time and effect involves a "boredom effect". This implies that after a certain amount of time, both in terms of length of individual training sessions and overall duration of the training, students are less attentive and do no better, or even worse, than students who spend less time in training.

A final explanation is that, perhaps, there is a saturation point for developing phonemic awareness. Perhaps with phonemic awareness training, like vitamin C in the human body, once a student has had enough, anything else is useless. Phonemic awareness skill is transferable to novel phonemes in like contexts (Byrne et al., 1991; Torgesen et al., 1992), thus, teaching students to segment or blend all of the phonemes is redundant. Perhaps after students have acquired the skill and practiced it for 20 minutes, practicing it an additional 15 minutes will not increase, and may decrease, their proficiency.

While all of the aforementioned justifications are plausible, none are particularly satisfying. Furthermore, there is no real indication in the data that the unusual results can be attributed to any of these explanations. Clearly, time as a continuous variable needs to be further investigated in relation to the size of the instructional group and to phonemic awareness effect sizes.

Group Size. Group Size was negatively and significantly correlated with effect size. As the size of the group increased, the effects of training decreased significantly. The relationship between Group Size and effect size remained significant when Total Minutes in Training was considered. These results were not surprising given the related research on class size.

Glass and Smith (1978, 1979, 1980) have conducted several meta-analyses of the research on reducing class size. They repeatedly found positive effects, both in academic and affective domains, for small classes over large classes. Slavin (1986) conducted a "best-evidence synthesis" of a portion of the research that explores the relationship between class size and achievement. He concluded that there were small, positive effects only in studies where class size was reduced by at least 40%.

While the research supporting small classes over large tends to show positive effects, class size is different from group size. Class size studies do not typically examine groups as small as those included in this study; what is considered a small class size (16-20) is considered a large phonemic awareness training group size. It would be inappropriate to generalize the positive results associated with the class-size reduction research to the correlation between Size of Training Group and effect sizes. On the other hand, it is reasonable to suggest that it is simply easier and more efficient to teach a group of five young children than it is to teach a group of fifteen.

Threats to the Validity of a Meta-Analysis

Missing Effect Sizes in Primary Studies

Fortunately, for this research synthesis, the original means and standard deviations for all included studies were either reported or available from the researcher. Thus, for the few times that researchers reported nonsignificant results without reporting numbers, effect sizes were still calculable.

The Lack of Statistical Independence Among Effect Sizes

Hedges (1994b) describes four types of threats to the assumption of independence among effect sizes; these were detailed in Chapter II. For all four of the threats to the assumption of independence among effect sizes described by Hedges, this study is guilty of assumption violation in the interest of pursuing revealing questions.

First, for this study's categorical analysis of Outcome Measures, groups of subjects were used more than once if they were administered more than one relevant dependent variable. Second, different experimental groups were compared to the same control group for the categorical analysis of Approach of Training. Third, studies were sometimes represented more than once in the synthesis. The results in Chapter IV consistently report when this happened. Finally, Byrne and Fielding-Barnsley (1991; 1995, Experiment #2) conducted two related studies which were included throughout this synthesis. One experiment (1995, Experiment #2) compared a new experimental group to the original control group. In further violation of the assumption of independence, several other researchers were represented more than once in the studies included in this review. See Appendix E.

Failure to Appropriately Weight Effect Sizes

This study utilized the Hedges and Olkin (1985) procedure in order to minimize the risk of inappropriately weighting studies. Interestingly, the studies with smaller sample sizes appeared to generate larger effect sizes, as indicated by correlations and scatterplots.

Publication Bias

For this research synthesis, dissertation abstracts and ERIC were accessed to locate any fugitive literature. However, once the reports were studied, it became apparent that most of the included dissertations had actually been published after their authors had taken university positions.

Publication bias is of limited concern for this study because the overall weighted mean is quite high (B. -T. Johnson, personal communication, March 26, 1999). There would have to be many overlooked studies with null or negative results for the effect size to move into the moderate or low category. A final guard against publication bias for phonemic awareness effect sizes is that the outliers identified during outlier diagnosis, which tended to be low, were included in the remaining moderator analyses.

The Lack of Statistical Power

Given that power is both a function of the number of studies reviewed (18) and the total number of subjects across studies (1202), the probability that this study appropriately reached significant conclusions is high.

For some analyses, the $\underline{d+}$ for a subgroup only represented a few effect sizes. For most of these analyses, the relationship between subgroups was not significant. This lack of power increases the likelihood of a Type II error, or the inability to reach a significant conclusion when the relationship is actually significant.

Instructional Implications

First, given the correlational relationship between phonemic awareness and learning to read (Bradley et al., 1983; Calfee et. al, 1973; Chall et al., 1963; Elkonin, 1973; Fox & Routh, 1975; Helfgott, 1976; Juel, 1988; Juel et al., 1986; Libermanet et al., 1977; Lomax et al., 1987; Lundberg et al., 1980; Rosner et al., 1971; Stanovich, Cunningham, & Feeman, 1984; Swank & Catts, 1991; Treiman & Baron, 1981; Tunmer, Herriman, & Nesdale, 1988; Tunmer et al., 1985; Zifcak, 1981) and the strong overall weighted mean effect for Phonemic Awareness produced by this study, teachers of young children should be intentionally teaching phonemic awareness. This study clearly indicates that if they do so, students will be more successful in phonemic awareness than untrained students.

While reading and spelling were not statistically aggregated in this synthesis, several studies included in this review examined the relationship between phonemic awareness training and reading achievement (Ayres, 1996; Ball et al., 1991; Blachman et al., 1994; Brady et al., 1994; Buys, 1992; Castle et al, 1994, Experiments 1 and 2; Cunningham, 1990; Davidson et al., 1994; Fox et al., 1984; O'Conner et al., 1995; Torgesen et al., 1992; Torgesen et al., 1996; Weiner, 1994) and spelling achievement (Ayres, 1996; Blachman et al., 1994; Brady et al., 1994; Castle et al., 1994, Experiments 1 and 2; Davidson et al., 1994; O'Conner et al., 1995; Torgesen et al., 1996). More specifically, qualitative examinations of these studies reveal that phonemic awareness training seems to enhance student reading performance on measures of decodable words (decodable pseudoword reading, decodable real words, and analogue tests). Connecting this information to the understanding that strength in phonemic awareness may lead to developing automaticity in word recognition which

facilitates success in comprehension (LaBerge & Samuels, 1985) provides yet another argument that it is worthwhile to teach phonemic awareness.

The results of the present review offer specific direction for instructional approach in phonemic awareness training. The findings suggest that the emphasis should be placed on segmenting and blending training. In examining the studies listed above which investigated phonemic awareness training as it relates to reading achievement and spelling achievement, it appeared that students trained in blending perform better on reading measures and students trained in segmenting perform better on spelling measures.

Many details of instruction remain unsolidified after this study. Incorporation of a sound to letter connection does not appear to support phonemic awareness acquisition. However, supplemental sound/letter incorporation does not seem to hinder phonemic awareness. As students must learn sound/letter correspondences in order to read, instruction in phoneme/grapheme relationships that is tangential to the auditory components of phonemic awareness training appears logical and meaningful. In all studies in this synthesis that incorporated graphemes, however, phonemes were first thoroughly introduced and manipulated without print (Ayres, 1993; Blachman et al., 1994; Castle et al, 1994, Experiments 1 and 2; O'Conner et al., 1995; Torgesen et al., 1996).

Because Goal of Training does not appear to influence learning, teachers should not fight the practical limitations of their classrooms by attempting to hold all students to a mastery criterion. On the other hand, Size of Training Group is related to developing Phonemic Awareness. In fact, training group size accounted for 28% of the variance among effect sizes. The studies that trained entire classes (Ayres, 1996; Brady et al., 1994, Byrne & Fielding-Barnesley, 1995, Experiment #2;) tended to

have the weakest phonemic awareness effect sizes (.41, .42, and .36 respectively).

However, the practical limitations of classrooms usually prohibit instruction in groups of 3-5, as was the case in the majority of training studies included in this synthesis (Ball & Blachman, 1991; Blachman, 1994; Castle et al, 1994, Experiments 1 and 2; Davidson & Jenkins, 1994; Fox & Routh, 1984; O'Conner et al., 1995; Rosner, 1971; Slocum et al., 1993; Torgesen et al., 1992; Torgesen & Davis, 1996; Weiner, 1994). Because Group Size was not examined in relation to reading achievement and spelling achievement, bending to classroom practicalities rather than declaring that phonemic awareness instruction must be with very small groups may be the sensible suggestion.

Because this study found that students trained for limited durations outperformed students trained at length, instructional groups should be watched to see when they demonstrate competence in phonemic awareness. Care should be taken not to inundate students with redundant and extensive practice without concern for whether their skill-level necessitates it.

Most importantly and most simply, teachers need to deliberately teach phonemic awareness. While instruction in phonemic awareness is likely to produce dramatic increases in phonemic awareness, it has not proven to be a spontaneous occurrence among children. Thus, schools must take specific initiative to train students to distinguish and manipulate the sounds within words.

Future Research

Given that there are hundreds of studies on phonemic awareness, it is difficult to imagine that there are holes in the research literature. However, this meta-analysis has highlighted a number of gaps in the research on phonemic awareness training as well as areas that have been persistently documented.

First, the question of whether phonemic awareness can be trained was clearly answered. While few doubted the answer before this study was produced, there is presently little room for speculation. In addition, training that emphasizes segmenting and blending training over instruction in one of these alone has been examined extensively in relation to phonemic awareness acquisition. However, training approach in relation to reading and spelling achievement was not investigated in this study and merits future statistical review.

When meta-analysts ask the same research questions as original researchers, the conclusions drawn by the meta-analysis tend to be particularly reliable. This meta-analysis, for example, examined whether training in segmenting and blending was more effective than training in segmenting or blending alone. There were several original researchers who asked the same research question. Consequently, the results for Approach of Training can be considered very reliable. On the other hand, study-level analysis was unavailable when examining the Goal of Training (mastery vs. nonmastery). Specifically, primary studies that directly compared a training group that was taught to mastery to a training group that was not taught to mastery would have strengthened the conclusions reached in this meta-analysis.

Primary research is needed, particularly in the following areas: the impact of supplementary sound/letter instruction; mastery-oriented training compared to lesson-oriented training; prekindergarten and first-grade studies; and large group instruction compared to small group instruction. Phonemic awareness training research conducted with intact classrooms and teachers, rather than with very small groups trained by a researcher, would enhance our understanding of the impact of phonemic awareness instruction in a "realistic" classroom setting. Most interestingly, the

relationship of duration of training, particularly in terms of minutes per training session and total minutes in training, merits further investigation.

Finally, there are many meta-analytic opportunities remaining within the current body of training studies. It would be interesting to include some of the studies excluded from this study on the basis that training was conducted in a language other than English, given that these studies are also examined as a class of their own. Meta-analytic possibilities include examination of the continuous relationship between time lapse from training to outcome measures, particularly for reading and spelling, and effect size. A meta-analyst may also want to look at more specific details of training such as the use of a puppet or other manipulatives.

For the sake of cumulative knowledge, it is very helpful if primary researchers report their results with the meta-analyst in mind. This means that researchers report the means and standard deviations for experimental and control groups. At some point a meta-analyst may want to explore ethnicity, intelligence quotient, first language, free/reduced lunch status (i.e., SES), and/or gender as moderating variables on the effects of training in phonemic awareness. Thus, it would be helpful for primary researchers to collect and report this information for each training and control group.

Another obvious meta-analytic opportunity is a synthesis of phonemic awareness training effects on reading and spelling achievement. Phonemic awareness is only interesting because it is correlated to reading ability. The million-dollar question is: Given that phonemic awareness can be trained with tremendous success, does this training result in improved reading and spelling abilities? Qualitative examinations of the studies included in this meta-analysis indicate that it does impact

word recognition. However, quantitative confirmation would extend our grasp of the accumulated literature.

Conclusions

This meta-analysis established the relationship between the independent variable and the dependent variable. It also investigated the degree to which moderator variables influenced this relationship.

First, the overall weighted mean describing the relationship between Phonemic Awareness Training and acquisition of Phonemic Awareness was high. Second, investigations of moderator variables demonstrated that several categorical moderators (Goal of Training, Context of Training, Grade-level at Training) tended not to influence the effects. On the other hand, continuous moderator variables associated with time (Minutes per Training Session, Total Minutes in Training) and Training Group Size demonstrated a negative and significant relationship with the effect sizes.

This review closed by articulating implications for classroom practice and by suggesting avenues for future research.

References

References marked with an asterisk indicate studies included in the meta-analysis. References marked with two asterisks contained two experiments included in the meta-analysis.

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. Review of Educational Research, 58, 151-179.

Adams, M. J. (1990). Beginning to read: Thinking and learning about print. Cambridge, MA: MIT Press.

Alexander, A. W., Anderson, H., Heilman, P. C., Voeller, K. S., & Torgesen, J. K. (1991). Phonological awareness training and remediation of analytic decoding deficits in a group of severe dyslexics. Annals of Dyslexia, 41, 193-206.

*Ayres, L. R. (1996). The efficacy of three training conditions on phonological awareness of kindergarten children and the longitudinal effect of each on later reading acquisition. Reading Research Quarterly, 30, 604-606.

Baldwin, S. R. & Vaughn, S. (1993). Response to Ridgeway, Dunston, & Qian: On methodological rigor: Has rigor mortis set in? Reading Research Quarterly, 28, 354-355.

*Ball, E. W. & Blachman, B. A. (1991). Does phoneme awareness training make a difference in early word recognition and developmental spelling? Reading Research Quarterly, 26, 49-66.

*Blachman, B. A., Ball, E. W., Black, R. S., & Tangel, D. M. (1994).

Kindergarten teachers develop phoneme awareness in low-income, inner-city classrooms: Does it make a difference? Reading and Writing: An Interdisciplinary Journal, 6, 1-18.

Blimling, G. S. (1988). Meta-analysis: A statistical method for integrating the results of empirical studies. Journal of College Student Development, 29, 543-549.

Bond, G. L., & Dykstra, R. (1967). The cooperative research program in first-grade reading instruction. Reading Research Quarterly, 2, 5-142.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read--a causal connection. Nature, 301, 208-225.

*Brady, S., Fowler, A., Stone, B., & Winbury, N. (1994). Training phonological awareness: A study with inner-city kindergarten children. Annals of Dyslexia, 44, 26-59.

Brown, J. & Brown L. (1987). Meta-analysis: Unraveling the mystery of research articles. The Volta Review, 89, 339-345.

*Buys, B. (1992). Three instructional strategies for teaching phonemic segmentation to kindergarten children. (Doctoral dissertation, State University of New York at Stony Brook, 1992). Dissertation Abstracts International, 55-11B, 5065-5143.

Bryant, P. E., MacLean, M., Bradley, L. L. & Crossland, J. (1990). Rhyme and alliteration, phoneme detection, and learning to read. Experimental Psychology, 26, 429-438.

Bryen, D. N. & Gerber, A. (1987). Metalinguistic abilities and reading: A focus on phonological awareness. Journal of Reading, Writing, and Learning DisabilitiesInternational, 3, 357-367.

*Byrne, B. & Fielding-Barnsley, R. (1991). Evaluation of a program to teach phonemic awareness to young children. Journal of Educational Psychology, 83, 451-455.

Byrne, B. & Fielding-Barnsley, R. (1993). Evaluation of a program to teach phonemic awareness to young children: A 1-year follow-up. Journal of Educational Psychology, 85, 104-111.

**Byrne, B. & Fielding-Barnsley, R. (1995). Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial. Journal of Educational Psychology, 87, 488-503.

Calfee, R. C., Lindamood, P., & Lindamood, C. (1973). Acoustic-phonetic skills and reading: Kindergarten through twelfth grade. Journal of Educational Psychology, 1973, 64, 293-298.

Carlberg, C. G. & Walberg, H. J. (1984). Techniques of research synthesis. The Journal of Special Education,18, 11-26.

Cary, L., & Verhaeghe, A. (1994). Promoting phonemic analysis ability among kindergarteners: Effects of different training programs. Reading and Writing: An Interdisciplinary Journal, 6, 251-78.

**Castle, J. M., Riach, J., & Nicholson, T. (1994). Getting off to a better start in reading and spelling: The effects of phonemic awareness instruction within a

whole language program. Journal of Educational Psychology, 86, 350-359.

Chall, J. S. (1983). Learning to read: The great debate (Rev. ed.) New York: McGraw-Hill Book Company.

Chall, J. S., Roswell, F. G., & Blumenthal, S. H. (1963). Auditory blending ability: A factor in success in beginning reading. The Reading Teacher, 17, 113-118.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Rev. ed.). New York: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Content, A., Kolinsky, R., Morais, J., & Bertelson, P. (1986). Phonetic segmentation in prereaders: Effect of corrective information. Journal of Experimental Child Psychology, 42, 49-72.

Cook, T. D. & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 48, 449-472.

Cooper, H. M. (1981). On the significance of effects and the effects of significance. Journal of Personality and Social Psychology, 41, 1013-1018.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

Cooper, H. M. & Arkin, R. M. (1981). On quantitative reviewing. Journal of Personality, 49, 225-230.

Cooper, H. M., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (pp. 3-14). Russell Sage Foundation: New York.

Cooper, H. M., & Hedges, L. V. (1994). Potentials and limitations of research synthesis. In H.M. Cooper & L.V. Hedges (Eds.), The Handbook of Research Synthesis (pp. 3-14). Russell Sage Foundation: New York.

Cunningham, A. E. (1990). Explicit versus implicit instruction in phonemic awareness. Journal of Experimental Child Psychology, 50, 429-444.

Davidson, M. & Jenkins, J. R. (1994). Effects of phonemic processes on word reading and spelling. Journal of Educational Research, 87, 148-157.

Davies, K. L. (1995). The effects on reading performance of matching student learning style/reading style preferences to complementary instructional methods of testing conditions: A meta-analysis. Unpublished doctoral dissertation,, University of Kansas, Lawrence.

Dunn, R. & Dunn, K. (1975). Educator's self-teaching guide to individualizing instructional programs. Reston, VA: Prentice-Hall.

Eagly, A. H. & Wood, W. (1994). Using research syntheses to plan future research. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (485-500). New York: Russell Sage Foundation.

Ehri, L. C., & Wilce, L. S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic? Reading Research Quarterly, 20,

163-179.

Elbro, C. (1996). Early linguistic abilities and reading development: A review

of a hypothesis. Reading and Writing: An Interdisciplinary Journal, 8, 453-485.

Elkonin, D. B. (1973). Methods of teaching reading. In J. Downing (Ed.), Comparative reading: Cross national studies of behavior and processing in reading and writing (pp.551-579). New York: McMillan.

Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.

Eysenck, H. J. (1984). Meta-analysis: An abuse of research integration. The Journal of Special Education, 18, 41-59.

Foorman, B. R., Francis, D. J., Novy, D.M., & Liberman, D. (1991) How letter-sound instruction mediates progress in first-grade reading and spelling. Journal of Educational Psychology, 83, 456-469.

Fox, B. & Routh, D. K. (1975). Analyzing spoken language into words, syllables and phonemes: A developmental study. Journal of Psycholinguistic Research, 4, 331-342.

Fox, B. & Routh, D. K. (1976). Phonemic analysis and synthesis as word-attack skills. Journal of Educational Psychology, 68, 70-74.

*Fox, B. & Routh, D. K. (1984). Phonemic analysis and synthesis as word attack skills: Revisited. Journal of Educational Psychology, 76, 1059-1064.

Gee, E. J. (1995). The effects of a whole language approach to reading

instruction on comprehension: A meta-analysis San Francisco: Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 384 003)

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher,5, 3-8.

Glass, G. V. (1978). Integrating findings: The meta-analysis of research. In L. S. Schulman (Ed.), Review of Research in Education, Itasca, Illinois: Peacock Publishers.

Glass, G. V., McGaw, B. & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills: Sage.

Glass, G. V. & Smith, M. L. (1978). Meta-analysis of research on the relationship of class size and achievement (Document No. OB-NIE-G-78-0103). San Francisco: Far West Laboratory for Educational Research and Development.

Glass, G. V. & Smith, M. L. (1979). Meta-analysis of research on the relationship of class size and achievement. Educational Evaluation and Policy Analysis,1, 2-16.

Glass, G. V. & Smith, M. L. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. American Educational Research Journal, 7, 419-433.

Greenhous, J. B. & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis

(383-398). new York: Russell Sage Foundation.

Griffith, P. L., Klesius, J. P. & Kromrey, J. D. (1992). The effect of phonemic awareness on the literacy development of first grade children in a traditional or a whole language classroom. Journal of Research in Childhood Education, 6, 35-92.

Hall, J. A., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (17-28). New York: Russell Sage Foundation.

Haller, E. P., Child, D. A., & Walberg, H. J. (1988). Can comprehension be taught? A quantitative synthesis of "metacognitive" studies. Educational Researcher, 17, 5-8.

Harckham, L. D. & Hagen, L. V. (1970). The effects of a phonics-oriented kindergarten program on auditory discrimination and reading readiness Minneapolis: Paper presented at the conference of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 040 834)

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490-499.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. American Psychologist, 42, 443-455.

Hedges, L. V. (1994a). The fixed effects model. In H. M. Cooper & L. V.

Hedges (Eds.), The Handbook of Research Synthesis (111-123). New York: Russell Sage Foundation.

Hedges, L. V. (1994b). Statistically analyzing effect sizes. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (111-123). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Helfgott, D. P. (1976). Phonemic segmentation and blending skills of kindergarten children: Implications for beginning reading acquisition. Contemporary Educational Research Psychology, 1, 157-169.

Holm, D. (1979). A simple sequential rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Houston, J. E. (ED.). (1995). Thesaurus of ERIC Descriptors (13th ed.). Phoenix: Oryx Press.

Hurford, D. P. & Sanders R. E. (1990). Assessment and remediation of a phonemic discrimination deficit in reading disabled second and fourth graders. Journal of Experimental Child Psychology, 50, 396-415.

Jackson, G. B. (1980). Methods of integrative reviews. Review of Educational Research, 50, 438-460.

Johnson, B. T. (1989). DSTAT: Software for the meta-analytic review of research literatures [computer software and manual]. Hillsdale, NJ: Lawrence

Erlbaum Associates, Inc.

Johnson, B. T. & Eagly, A. H. (in press). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), Handbook of Research Methods in Social Psychology. London: Cambridge University Press.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. Journal of Educational Psychology, 80, 437-447.

Juel, C., Griffith, P. L., & Gough, P. B. (1986). The acquisition of literacy: A longitudinal study of children in first and second grade. Journal of Educational Psychology, 78, 243-255.

Kavale, K. A. (1988). Using meta-analysis to answer the question: What are the important, manipulable influences on school learning? School Psychology Review, 17, 644-650.

Klesius, J. P., Griffith, P. L., & Zielonka, P. (1991). A whole language and traditional instruction comparison: Overall effectiveness of the alphabetic principle. Reading Research and Instruction, 30, 47-61.

Kraiger, K. (1985). On learning from the past: A meta-analytic fable. Personnel Psychology, 38, 799-802.

Kulik, J. A. & Kulik, C. L. C. (1989). Meta-analysis in education. International Journal of Educational Research, 13, 221-340.

Kulik, C. C., Kulik, J. A., & Banger-Drowns, R. L. (1990). Effectiveness of

mastery learning programs: A meta-analysis. Review of Educational Research, 60, 265-299.

LaBerge, D. & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology, 6, 293-323.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fischer, F. W. (1977). Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. Scarborough (Eds.), Toward a psychology of reading: The proceedings of the CUNY conference (pp.207-225). Hillsdale, NJ: Earlbaum.

Lie, A. (1991). Effects of a training program for stimulating skills in word analysis in first-grade children. Reading Research Quarterly, 26, 234-250.

Light, R. J. & Pillemer, D. B. (1982). Numbers and narrative: Combining their strengths in research reviews. Harvard Educational Review, 52, 1-26.

Light, R. J. & Pillemer, D. B. (1984). Summing up: The science of reviewing research. Cambridge, MA: Harvard University Press.

Light, R. J. & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 42, 429-471.

Lindamood Auditory Conceptualization Test, (1979). New York Teaching Resources, The New York Times.

Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (111-123). New York: Russell Sage Foundation.

Lomax, R. G., & McGee, L. M. (1987). Young children's concepts about

print and reading: Toward a model of word reading acquisition. Reading Research Quarterly, 22, 237-256.

Lundberg, I., Olofsson, A., & Wall, S. (1980). Reading and spelling skills in the first school years predicted from phonemic awareness skills in kindergarten. Scandinavian Journal of Psychology, 21, 159-173.

Manke, M. P. & Erwin, T. D. (1988). Meta-analysis: Summarizing student development literature. Journal of College Student Development, 29, 549-552.

Mathes, P. G., & Fuchs, L. S. (1991). The efficacy of peer tutoring in reading for students with disabilities: A best-evidence synthesis Washington, DC: Special Education Programs. (ERIC Document Reproduction Service No. ED 344 352)

Matt, G. E. & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (111-123). New York: Russell Sage Foundation.

McNeil, J. D. & Coleman, J. C. (1967). Auditory discrimination training in the development of word analysis skills (Project No. 5-0503). Los Angeles: U.S. Department of Health, Education, and Welfare. (ERIC Document Reproduction Service No. ED 018 344)

Miller, N. & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research

Synthesis (111-123). New York: Russell Sage Foundation.

Minus, M. E. (1992). The relationship of phonemic awareness to reading level and the effects of phonemic awareness instruction on the decoding skills of adult disabled readers San Antonio: Paper presented at the National Reading Conference. (ERIC Document Reproduction Service No. ED 352 641)

Murray, B. A. (1996). Developing phoneme awareness through alphabet books. Reading and Writing: An Interdisciplinary Journal, 8, 307-322.

Naslund, J. C. & Schneider, W. (1996). Kindergarten letter knowledge, phonological skills, and memory processes: Relative Effects on Early Literacy. Journal of Experimental Child Psychology, 62, 30-59.

Neilsen, L. (1993). Response to Ridgeway, Dunston, & Qian: Authoring the questions: Research as an ethical enterprise. Reading Research Quarterly, 28, 350-353.

Nicholson, T. (1996). The great debate. Recent research on reading (and a brief look at the "Reading Debate". Nga-Kete-Korero: Journal of the Adult Reading & Learning Assistance Federation, 4, 12-17.

O'Conner, R. E., Jenkins, J. R., Leicester, N., & Slocum, T. A. (1992). Teaching phonemic awareness to young children with disabilities: Blending, segmenting, and rhyming San Francisco: paper presented at the annual conference of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 349 761)

*O'Conner, R. E., Jenkins, J. R., & Slocum, T. A. (1995). Transfer among

phonological tasks in kindergarten: Essential instructional content. Journal of Educational Psychology, 87, 202-217.

O'Conner, R. E. & Notari-Syverson, A. (1995). Ladders to Literacy: The effects of teacher-led phonological activities for kindergarten children with and without disabilities San Francisco: Paper presented at the annual conference of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 385 378)

Olofsson, A. & Lundberg, I. (1983). Can phonemic awareness be trained in kindergarten? Scandinavian Journal of Psychology, 24, 35-44.

Orwin, R. G. (1994). Evaluating coding decisions. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (111-123). New York: Russell Sage Foundation.

Perfetti, C. A. (1984). Reading acquisition and beyond: Decoding includes cognition. American Journal of Education, 93, 40-60.

Pflaum, S. W. (1982). Synthesizing research in reading. Reading Psychology, 3, 325-337.

Pillemer, D. B. (1984). Conceptual issues in research synthesis. The Journal of Special Education, 18, 27-40.

Read, C., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic reading. Cognition, 24, 31-44.

Readence, J. E. & Moore, D. W. (1981). A meta-analytic review of the effect of adjunct pictures on reading comprehension. Psychology in the Schools, 18, 218-224.

Ribowsky, H. (1985). The effects of a code emphasis approach and a whole language approach upon emergent literacy of kindergarteners (ERIC Document Reproduction Service No. ED 269 720)

Rosenshine, B. & Meister, C. (1994). Reciprocal teaching: A review of the research. Review of Educational Research, 64, 479-530.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R. & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166-169.

*Rosner, J. (1971). Phonic analysis training and beginning reading skills Washington, DC: Paper presented at the annual meetings of the American Psychological Association. (ERIC Document Reproduction Service No. ED 059 029)

Rosner, J. (1974). Auditory analysis training with prereaders. Reading Teacher, 27, 379-384.

Rosner, J. & Simon, D. (1971). The Auditory Analysis Test: An initial report. Journal of Learning Disabilities, 4, 384-392.

Shadish, W. R., & Haddock, C. K. (1994). Combining effect sizes for categorical data. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (261-281). New York: Russell Sage Foundation.

Shavelson, R. J. (1988). Statistical Reasoning for the Behavioral Sciences (2nd Ed.). Boston: Allyn and Bacon, Inc.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used? Educational Researcher, 13, 6-15.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. Educational Researcher, 15, 5-11.

Slavin, R. E. (1987a). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. Review of Educational Research, 57, 293-336.

Slavin, R. E. (1987b). Mastery Learning: A best evidence synthesis. Review of Educational Research, 57, 175-213.

Slavin, R. E. (1990). Mastery learning reconsidered. Review of Educational Research, 60, 300-302.

*Slocum, T. A., O'Conner, R. E., & Jenkins, J. R. (1993). Transfer among phonological manipulation skills. Journal of Educational Psychology, 85, 618-630.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Smith, N. B. (1974). American reading instruction. Newark, DE: International Reading Association.

Smith, S. B. (1995). Phonological awareness: Curricular and instructional implications for diverse learners (Technical Report No. 22). Eugene, OR: National Center To Improve the Tools of Educators. (ERIC Document Reproduction Service No. 386 869)

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). Preventing Reading Difficulties in Young Children. Washington, DC: National Academy Press.

Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. Journal of Educational Psychology, 84, 353-363.

Speer, O. B., & Lamb, G. D. (1976). First grade reading ability and fluency in naming verbal symbols. Reading Teacher, 26, 572-576.

Spedding, S. & Chan L. K. S. (1993). Metacognition, word identification, and reading competence. Contemporary Educational Psychology, 18, 91-100.

Stahl, S. A. & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. Review of Educational Research, 56, 72-110.

Stahl, S. A., McKenna, M. C., & Panucco, J. R. (1993). The effects of whole language instruction: An update and reappraisal Charleston: paper presented at the annual meeting of the National Reading Conference. (ERIC Document Reproduction Service No. ED 364 830)

Stahl, S. A. & Miller, P. D. (1989). Whole language and language experience approaches for beginning reading: A quantitative research synthesis. Review of

Educational Research, 59, 87-116.

Stahl, S. A. & Murray, B. A. (1994). Defining phonological awareness and its

relationship to early reading. Journal of Educational Psychology, 86, 221-234.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly, 21, 360-407.

Stanovich, K. E., Cunningham, A. E., & Cramer, B. B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. Journal of Experimental Child Psychology, 38, 175-190.

Stanovich, K. E., Cunningham, A. E., & Feeman, J. (1984). Intelligence, cognitive skills, and early reading progress. Reading Research Quarterly, 19, 120-139.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (125-138). New York: Russell Sage Foundation.

Swank, L. K., & Catts, H. W. (1991). Phonological awareness Las Vegas: Paper presented at the annual meeting of the International Reading Association. (ERIC Document Reproduction Services No. ED 335 663)

Swanson, H. L. (1996). Meta-analysis, replication, social skills, and learning disabilities. The Journal of Special Education, 30, 213-221.

*Torgesen, J. K., Morgan, S. T. & Davis, C. (1992). Effects of two types of

phonological awareness training on word learning in kindergarten children. Journal of Educational Psychology, 84, 364-370.

*Torgesen, J. K. & Davis, C. (1996). Individual difference variables that predict response to training in phonological awareness. Journal of Experimental Child Psychology, 63, 1-21.

Treiman, R. & Baron, J. (1981). Semental analysis ability: Development and relation to reading ability, In T. G. Waller & G. E. MacKinnon (Eds.), Reading Research: Advances in Theory and Practice (pp.159-198). New york: Academic Press.

Treiman, R. & Baron, J. (1983). Phonemic analysis training helps children benefit from spelling-sound rules. Memory and Cognition, 11, 382-389.

Troia, G. A. (1999). Phonological awareness intervention research: A critical review of the experimental methodology. Reading Research Quarterly, 34, 28-52.

Tunmer, W. E., & Nesdale, A. R. (1985). Phonemic segmentation skill and beginning reading. Journal of Educational Psychology, 77, 417-427.

Tunmer, W. E. & Rohl, M. (1991). Phonological awareness and reading acquisition. In D. J. Sawyer, B. J. Fox (Eds.), Phonological awareness in reading: The evolution of current perspectives (pp. 1-30). New York: Springer-Verlag.

Tunmer, W. E., Herriman, M. L., & Nesdale, A. R. (1988). Metalinguistic abilities and beginning reading. Reading Research Quarterly, 23, 134-158.

*Weiner, S. (1994). Effects of phonemic awareness training on low- and

middle-achieving first graders' phonemic awareness and reading ability. Journal of Reading Behavior, 26, 277-300.

White, H. D. (1994). Scientific communication and literature retrieval. In H. M. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (41-55). New York: Russell Sage Foundation.

Williams, J. P. (1980). Teaching decoding with an emphasis on phoneme analysis and blending. Journal of Educational Psychology, 72, 1-15.

Winsor, P. J. T. (1990). Developing phonemic awareness: Knowledge and practice in holistic instruction. Miami: Paper presented at the annual meeting of the National Reading Conference. (ERIC Document Reproduction Service No. ED329933)

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. Reading Research Quarterly, 28, 159-177.

Zifcak, M. (1981). Phonological awareness and reading acquisition. Contemporary Educational Psychology, 6, 117-126.

Zhurova, L. Y. (1973). The development of analysis of words into their sounds by preschool children. In C.A Ferguson and D.I. Slobin (Eds.). Studies of Child Language Development. New York: Holt, Rinehart, and Winston.

Appendix A

Coding Form

I. General Information

Article Title_____

Article ID#_____Author_____Year of Publication_____

Journal___ Dissertation/Thesis___

Other_____

Journal Title_____

II. Quality

1. Source of data: Published ___(1) Unpublished___(2)

2. Sampling: Some attempt at representation___(1) Convenience___(0)

3. Sample size: Large (40+in C &E each)___(2) Medium(20-39 in C&E each)___(1)
   Small___(0)

4. Assignment to groups:

   Control and experimental groups matched? much (6 or more)___(2); some

   (1-5)___(1); none___(0)

   age___ gender___ grade___ IQ___ SES___ race___ ABC___ PA___

   other pretests___ classroom___

Random assignment?____(3) Random assignment within controlled variables?___(4)

5. Groups treated equally? yes___(1) no___(0)

6. Same trainer for all groups? yes___(1) no___(0) unknown___(0)

7. Checks on trainer fidelity to treatment? yes___(1) no___(0)

8. Person administering post-tests is blind to group assignment? yes___(1), no___(0)

   unknown___(0)

9. Experimenter trained: control group ___(0), experimental group___(0),

   both___(1), neither___(2), unknown___(0).

Total Score for Study ___/15

Score assigned by Troia (1999): _____

III. Subjects: Total subjects:_____

Experimental Group 1:

General: # subjects___ **Mean Age** (in mos.): 48-59 (4 to 5 yrs.)___ 60-65 (5 to 5.5 yrs.)___ 66-71(5.5 to6 yrs.)___ 72-77 (6 to 6.5 yrs.)___ 78-83 (6.5 to 7 yrs.) **Grade:**____

Control Group:

General: # subjects___ **Mean Age** (in mos.): 48-59 (4 to 5 yrs.)___ 60-65 (5 to 5.5 yrs.)___ 66-71(5.5 to6 yrs.)___ 72-77 (6 to 6.5 yrs.)___ 78-83 (6.5 to 7 yrs.) **Grade:**____

Experimental Group 2:

General: # subjects___ **Mean Age** (in mos.): 48-59 (4 to 5 yrs.)___ 60-65 (5 to 5.5 yrs.)___ 66-71(5.5 to6 yrs.)___ 72-77 (6 to 6.5 yrs.)___ 78-83 (6.5 to 7 yrs.) **Grade:**____

Experimental Group 3:

General: # subjects___ **Mean Age** (in mos.): 48-59 (4 to 5 yrs.)___ 60-65 (5 to 5.5 yrs.)___ 66-71(5.5 to6 yrs.)___ 72-77 (6 to 6.5 yrs.)___ 78-83 (6.5 to 7 yrs.) **Grade:**____

IV. Training

Experimental Group 1:

Categorical Moderator Variables:

Approach:

___SEGMENTATION (Analysis): Counting___ Phonemes___ Comparing___ Free Segmentation___ Tapping___ O/R___ Manipulatives___

___BLENDING (Synthesis): O/R___ Phonemes___

___DELETION: Initial___ Final___ Medial___

___ADDITION: Initial___ Final___ Medial___

___COMBINATION: Substitution___ Reversal___

Goal: Mastery Criterion___ No Mastery Criterion_

Context: Sound/Letter Associations Taught___ No Sound/Letter Associations___

Continuous Moderator Variables:

Group size: _____

Duration of training: _____ minutes per training session

Training lasted _____ weeks. Total minutes in training _____

Frequency of training: Subjects met with a trainer _____ times per week.

Comments:

Control Group

Continuous Moderator Variables: No Training___

Group size: _____

Duration of training: _____ minutes per training session

Training lasted _____ weeks. Total minutes in training _____

Frequency of training: Subjects met with a trainer _____ times per week.

Trainer:_____

Description of Training:

Experimental Group 2

Categorical Moderator Variables:

Approach:

___SEGMENTATION (Analysis): Counting___ Phonemes___ Comparing___ Free Segmentation___ Tapping___ O/R___ Manipulatives___

___BLENDING (Synthesis): O/R___ Phonemes___

___DELETION: Initial___ Final___ Medial___

___ADDITION: Initial___ Final___ Medial___

___COMBINATION: Substitution___ Reversal___

Goal: Mastery Criterion___ No Mastery Criterion_

Context: Sound/Letter Associations Taught___ No Sound/Letter Associations___

Continuous Moderator Variables:

Group size: _____

Duration of training: _____ minutes per training session

Training lasted _____ weeks. Total minutes in training _____

Frequency of training: Subjects met with a trainer _____ times per week.

Comments:

Experimental Group 3

Categorical Moderator Variables:

Approach:

___SEGMENTATION (Analysis): Counting___ Phonemes___ Comparing___ Free Segmentation___ Tapping___ O/R___ Manipulatives___

___BLENDING (Synthesis): O/R___ Phonemes___

___DELETION: Initial___ Final___ Medial___

___ADDITION: Initial___ Final___ Medial___

___COMBINATION: Substitution___ Reversal___

Goal: Mastery Criterion___ No Mastery Criterion_

Context: Sound/Letter Associations Taught___ No Sound/Letter Associations___

Continuous Moderator Variables:

Group size: _____

Duration of training: _____ minutes per training session

Training lasted _____ weeks. Total minutes in training _____

Frequency of training: Subjects met with a trainer _____ times per week.

Comments:

V. Outcome Measures

1.___SEGMENTATION (Analysis): Name of Test_____

2.___BLENDING (Synthesis): Name of Test_____

3.___DELETION: Name of Test_____

4.___COMBINATION MEASURES: Name of Test_____

(substitution, reversal, Other_____)

Appendix B

Coding Information Sheet

II. Quality

1. Source of data: Published ___(1) Unpublished___(2)

Journal articles are considered published. Dissertations, conference presentations, and government research studies are unpublished.

2. Sampling: Some attempt at representation___(1) Convenience___(0)

In order to mark "some attempt at representation", the words "random sampling" need to be used in the description of the sampling procedures. If the words "random sampling" are not used, then it is a convenience sample.

3. Sample size: Large (40+in C &E each)___(2) Medium(20-39 in C&E each)___(1)

Small___(0)

C & E refer to control and experimental groups.

4. Assignment to groups:

Control and experimental groups matched? much (6 or more)___(2); some

(1-5)___(1); none___(0)

age___ gender___ grade___ IQ___ SES___ race___ ABC___ PA___

other pretests___ classroom___

For every student placed in the experimental group a student which had the same age/gender/grade/IQ, etc. was placed in the control or other experimental groups.

Random assignment?____(3)

Random assignment within controlled variables?___(4)

In order to check "random assignment" the study must use those exact words. For example, a study might say, "The 75 subjects were randomly assigned to either the control or the experimental group." Note: random sampling (#2) and random assignment to groups are two different things.

Random assignment within controlled variables means that students were randomly assigned but certain conditions were met. For example, the low, medium, and high students were identified and then evenly and randomly distributed among the groups. Basically, the number of subjects representing a particular variable is controlled within a group. If a group is to have half boys and half girls, the experimenter may randomly assign a certain number of boys and a certain number of girls to each group.

5. Groups treated equally? yes___(1) no___(0)

Generally, this item refers to whether the control group is treated similarly to the experimental group. Does the control group receive treatment that is comparable to the experimental group in terms of group size, duration of training, etc. If the control group receives no training, then the groups are not treated equally. However, in some situations, one experimental group can actually be analyzed as the control group. In this case, the "no training" group would not be included in this analysis.

6. Same trainer for all groups? yes___(1) no___(0) unknown___(0)

7. Checks on trainer fidelity to treatment? yes___(1) no___(0)

Did the experimenter take measures to ensure that the trainer implemented the training the way the experimenter intended. (i.e., scripted program, surprise observations, random videotaping, etc.)

8. Person administering post-tests is blind to group assignment? yes___(1), no___(0) unknown___(0)

Does the person assessing the students at the end of the training know which children were in the experimental groups and which were in the control group?

9. Experimenter trained: control group ___(0), experimental group___(0), both___(1), neither___(2), unknown___(0).

IV. Training

Categorical Moderator Variables:

Approach:

___SEGMENTATION (Analysis): Counting___ Phonemes___ Comparing___ Free Segmentation___ Tapping___ O/R___ Manipulatives___

Segmentation is any training task requiring subjects to break words into parts. A phoneme is the smallest sound you can produce and may be represented by more than one letter. /Sh/ is one phoneme represented by two letters.

Segmentation may also be referred to as analysis. These tasks include:

***Counting**-subjects count the phonemes in a word (ex. dog=3 phonemes; wish=3 phonemes).*

***Phonemes**-subjects are simply required to break words into their smallest parts (phonemes). Ex. "What is the beginning sound in the word fish?" /f/; "Say the sounds in cat. /c/-/a/-/t/"*

***Comparing**-subjects are asked to compare words. For example, the experimenter may say a word and ask the student to point to a picture of something that starts the same way. The subject may be asked to look at three pictures or listen to three words and tell which does not have the same ending sound.*

***Free Segmentation**-subjects segment a word without specific guidelines about where to segment. For example, student may be told to "say part of the word" or "say a little bit of the word". The word milk could be segmented as /m/-/ilk/, /mi/-/lk/; /mil/-/k/, etc.*

Tapping-*Students are asked to tap the number of phonemes in a word. For the word "dog" a student would tap three times.*

O/R (Onset-Rime)-*subjects are asked to segment words into their initial consonant or consonant blend and the vowel portion. Ex: m-ilk, tr-ee; f-ish; str-eet; j-ump, etc.*

Manipulatives-*Three specific tasks involve manipulatives: Elkonin Boxes, Say-it/Move-it, and Lindamood. Any task that requires students to use blocks, discs, tokens, or some other manipulatives to represent the sounds in a word would fit into this category.*

___BLENDING (Synthesis): O/R___ Phonemes___

When students are given phonemes or a combination of phonemes in sequence and then asked to produce a corresponding word, they are blending. Blending is the reverse process of segmenting. Blending tasks include:

O/R (Onset/Rime)-*Subjects are given the onset /r/ and the rime /un/ of a word and asked to produce the word /run/.*

Phonemes-*Subjects are given the individual phonemes /r/-/u/-/n/ and are asked to produce the word /run/.*

___DELETION: Initial___ Final___ Medial___

Students are asked to repeat a word, omitting a particular phoneme.

Initial: *"Say the word fan without the /f/."*

Medial: *"Say the word monkey without the /k/."*

Final: *"Say the word cart without the /t/."*

___ADDITION: Initial___ Final___ Medial___

Students are asked to repeat a word, adding a particular phoneme.

Initial: *"Say the word an with /f/ at the beginning."*

Medial: *"Say the word money with a /k/ in the middle."*

Final: *"Say the word car with a /t/ at the end."*

___COMBINATION: Substitution___ Reversal___

These are tasks that draw on a combination of the aforementioned tasks.

Substitution: *Involves deleting and adding phonemes. For example, "What would you have if you changed the /y/ in yellow to an /m/?*

Reversal: *Involves segmenting and blending as students reorder the word. For example, "What word would you have if you said the word "tan" backward.*

Continuous Moderator Variables:

Goal: Mastery Criterion___ No Mastery Criterion___

Mastery*: Students had to demonstrate competence with the taught task before they were given the post-test. Discussion that refers to some children needing more instruction than others would indicate teaching to mastery. If a range of time in training is given, it could be a clue that students were taught to mastery (i.e., "Training sessions lasted 10-18 minutes depending on how quickly children learned the task.").*

Non-Mastery: *The opposite of mastery. If all students received the same amount of training, they were not taught to mastery.*

Context: Sound/Letter Associations Taught___ No Sound/Letter Associations___

Sound/Letter Associations Taught: *While the focus of all training is phonemic awareness, some training programs introduce graphemes in the last few training sessions. For example, students might initially use discs to represent phonemes (Say-it/Move-it) and then later the actual letters are written on the discs. Substitution tasks may be demonstrated using magnetic letters.*

V. Outcome Measures

The definitions detailed in Section IV apply to phonemic awareness assessments.

Phonemic Awareness

___SEGMENTATION (Analysis): Name of Test_____

___BLENDING (Synthesis): Name of Test_____

___DELETION: Name of Test_____

___COMBINATION: Name of Test_____

Appendix C

Coding of Categorical Variables

The categorical definitions listed below are for interpreting the following table:

approach of training: 1=segmenting; 2=blending; 3=segmenting & blending; 4=combination;

goal of training: 1=mastery oriented; 2=lesson oriented;

context of training: 1=supplementary letter names incorporated; 2= no letter names;

grade of subjects at training: 1=pre-kindergarten; 2=kindergarten; 3=first grade;

The outcome measures associated with the studies is not presented because the variation within studies makes them difficult to report.

Coding of Categorical Variables

Study #	Category	Goal	Grade	Context
1	1	2	1	2
2	1	2	1	2
3	4	2	2	1
4	4	2	2	1
5	V	1	2	2
6	4	1	3	2
7	V	1	1	2
8	V	1	2	2
9	4	2	3	2
10	4	1	2	2
11	V	1	2	2
12	3	2	V	2
13	1	2	2	1
14	1	2	2	1
15	3	1	2	1
16	1	2	2	1
17	1	2	2	2
18	4	2	2	1

Note. "V" indicates variation within the study.

Appendix D

Coding of Continuous Variables

Study #	Quality	Index	Minutes	Total Minutes	Group Size	Frequency
1	8	19	27.5	330	5	1
2	3	19	27.5	330	20	1.3
3	6	20	20	400	5	2
4	8	19	20	300	3.5	1
5	9	na	10	150	4	5
6	8	na	17.5	647.5	8	5
7	9	15	10	51	1	5
8	9	14	20	450	4	3
9	5	26	25	300	4.5	2
10	6	29	20	1080	na	3
11	9	32	15	337.5	5.5	4.5
12	9	21	17.5	350	4.5	2
13	9	na	17.5	770	4.5	4
14	8	na	20	V	V	5
15	11	29	20	960	4	4
16	10	23	20	560	5	4
17	9	na	13.5	432	20	4
18	8	30	15	300	4	2

Note. "V" indicates that the moderator varied within the study.

Appendix E

Study Identification Numbers

Study #	Authors	Publication Date
1	Byrne & Fielding-Barnsley	1991
2	Byrne & Fielding-Barnsley	1995 (Experiment #2)
3	Castle, Riach, & Nicholson	1994 (Experiment #1)
4	Castle, Riach, & Nicholson	1994 (Experiment #2)
5	Davidson & Jenkins	1994
6	Rosner	1971
7	Slocum, O'Conner, & Jenkins	1993
8	Torgesen, Morgan, & Davis	1992
9	Weiner	1994
10	Brady, Fowler, Stone, & Windbury	1994
11	Fox & Routh	1984
12	Cunningham	1990
13	Blachman, Ball, Black, & Tangel	1994
14	Ayres	1993
15	Torgesen & Davis	1996
16	Ball & Blachman	1991
17	Buys	1992
18	O'Conner, Jenkins, & Slocum	1995