Education Policy Analysis Archives

Volume 11 Number 15 **May 8, 2003** **ISSN 1068-2341**

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Teaching Children to Read: The Fragile Link Between Science and Federal Education Policy

Gregory Camilli
Sadako Vargas
Michele Yurecko

National Institute for Early Education Research
and
Rutgers University

Citation: Camilli, G., Vargas, S., and Yurecko, M. (May 8, 2003). *Teaching Children to Read*: The fragile link between science and federal education policy. *Education Policy Analysis Archives, 11*(15). Retrieved [date] from http://epaa.asu.edu/epaa/v11n15/.

Abstract
Teaching Children to Read (TCR) has stirred much controversy among reading experts regarding the efficacy of phonics instruction. This report, which was conducted by the National Reading Panel (NRP), has also played an important role in subsequent federal policy regarding reading instruction. Using *meta-analysis*, the NRP found that systematic phonics instruction was more effective than alternatives in teaching children to read. In the present

study, the findings and procedures leading to TCR were examined. We concluded that the methodology and procedures in TCR were not adequate for synthesizing the research literature on phonics instruction. Moreover, we estimated a smaller though still substantial effect (d = .24) for systematic phonics, but we also found an effect for systematic language activities (d = .29) and tutoring (d = .40). Systematic phonics instruction when combined with language activities and individual tutoring may *triple* the effect of phonics alone. As federal policies are formulated around early literacy curricula and instruction, these findings indicate that phonics, as one aspect of the complex reading process, should not be over-emphasized.

The data files that serve as the basis of this article are available for download.

Executive Summary

In 1997 the U.S. Congress directed the Director of the National Institute of Child Health and Human Development (NICHD), in consultation with the Secretary of Education, to establish a national panel on research in early reading development. The panel, now known as the National Reading Panel (NRP), was charged with conducting a thorough study of the research, determining what research findings were suitable for classroom application, and recommending methods of dissemination. Six areas of reading were eventually examined, and an influential report was released in December 2000. This report, *Teaching Children to Read*, has stirred much controversy among reading experts, and both critics and supporters have been highly visible in national-level venues. Without question, the report has played an important role in subsequent federal policy regarding reading instruction.

One of the six areas of reading research examined by the NRP was phonics instruction. According to the NRP:

> An essential part of the process for beginners involves learning the alphabetic system, that is, letter-sound correspondences and spelling patterns, and learning how to apply this knowledge in their reading. Systematic phonics instruction is a way of teaching reading that stresses the acquisition of letter-sound correspondences and their use to read and spell words…. (NRP, 2000b, p. 2-89).

Using a research methodology known as *meta-analysis*, the NRP identified 38 experimental and quasi-experimental—meaning a reasonably close approximation to experimental—research studies on phonics instruction. (A meta-analysis can be thought of as a quantitative literature review.) Based on a statistical "averaging" of the outcomes from these 38 studies, the NRP concluded that their findings "provided solid support" for the conclusion that systematic phonics instruction is more effective than alternatives in teaching children to read. Altogether, eleven conclusions were offered regarding the efficacy of phonics instruction, but the above finding is of prime importance.

In their deliberations on research findings, the NRP clearly recognized the ultimate need for instructional decisions to be based on the best empirical evidence and methods of analysis. The NRP recounted that one theme "expressed repeatedly," at a series of five regional public hearings held prior to its work, was the importance of high standards for choosing evidence about what works in reading instruction. The NRP interpreted this to mean that experimental and quasi-experimental studies were most likely to contain reliable, valid, and replicable findings. However, two aspects of the scientific method are important and should be distinguished. The review process, i.e., meta-analysis, is a set of procedures for distilling conclusions and generalizations from research studies. In contrast, the "standards of scientific evidence"—which led the NRP to focus on experimental studies—determine what evidence will be included in the meta-analytic process.

For the purposes of this review, we were primarily concerned with the former aspect, that is, the research review process. Most currently available reviews of the NRP's study have focused on the *interpretation* of the results for phonics instruction while assuming the basic correctness of the measurement and analytic procedures. We did not make such assumptions; rather, we designed an independent study in an attempt to reconstruct the NRP's central findings. As in other types of scientific investigation, replicability is a key criterion for judging the credibility of the NRP meta-analysis, and consequently how seriously we should consider applying its findings.

We began with the *same* 38 studies analyzed by the NRP, but in the course of our analysis, we deleted one study and added three. We then devised alternative plans for extracting and analyzing data from 40 studies (38 – 1 + 3 = 40). Based on these analyses, conclusions were drawn and interpretations made regarding the efficacy of phonics instruction. Though some of the methodological steps taken by the NRP analysts were retraced, our goal was to verify whether an independent team of researchers would arrive at conclusions consistent with those in the NRP report. We did not examine how the original 38 studies were chosen. It would have been useful to examine the full range of the NRP's procedures and findings, including study selection, but this task would have required resources well beyond our means.

In our analyses, we found that programs using systematic phonics instruction outperformed programs using less systematic phonics with $d = .24$. Though this effect is statistically significant, it was substantially smaller than the estimate of the NRP at $d = .41$. (Roughly speaking, $d = 0$ means no effect; $d = .5$ is moderate; and $d = 1.0$ is large.) The systematic phonics effect, moreover, was smaller than the effect for individual tutoring ($d = .40$). Students receiving tutoring had one-to-one instruction as opposed to instruction in small groups or classes. We also found that students who received systematic language activities did better ($d = .29$). This effect is comparable to that of systematic phonics instruction. In addition, standardized tests tended to give larger effects than locally developed instruments (*d* = .19). Overall, we concluded that there is reason to believe that these effects are additive. Systematic phonics instruction when combined with language activities and individual tutoring may *triple* the effect of phonics alone.

Though language activities were included in over 30% of the treatment conditions in the 38 studies, the NRP analysts missed the language effect for one simple reason: they didn't look for it. In our opinion, an approach that recognizes the complexity of reading instruction has the potential to improve the estimates of average effect sizes in *all* substantive areas that the NRP examined including: phonemic awareness instruction; fluency; comprehension; vocabulary instruction; text comprehension instruction; teacher preparation and comprehension; strategies instruction; teacher education and reading instruction; and computer technology and reading instruction. To obtain more accurate estimates of the full range of variables that influence reading, analyses would also benefit from, and indeed may require, a substantially larger sample of studies. In this effort, researchers with substantive, methodological, and classroom experience—as well as time and resources—are necessary to find studies, and to propose and test alternative design strategies. While we applaud the NRP for taking the challenging and difficult first steps in summarizing the extant knowledge on reading instruction, it is clear that substantial resources will be required for completing this essential work.

If the NRP results are taken to mean that effective instruction in reading should focus on phonics to the exclusion of other curricular activities, instructional policies are likely to be misdirected. This interpretation of the data results from a design in which simultaneous influences on reading interventions were not adequately coded and analyzed. In particular, early literacy policies are a timely concern, especially as they are interpreted and applied in the federal Early Reading First Program. Program administrators and teachers need to understand that while scientifically-based reading research supports the role of phonics instruction, it also supports a strong language approach that provides individualized instruction. As federal policies are formulated around early literacy curricula and instruction, it is important not to over-emphasize one aspect of a complex process. Fletcher and Lyon (1998) wrote "a targeted skill cannot be learned without opportunities for practice and application." With this common sense observation in mind, it is not surprising that the research shows a balance of systematic phonics, tutoring, and language activities is best for teaching children to read.

Introduction

In 1997 the U.S. Congress directed the Director of the National Institute of Child Health and Human Development (NICHD), in consultation with the Secretary of Education, to establish a national panel on research in early reading development. The panel, which is now known as the National Reading Panel (NRP), was charged with conducting a thorough study of the research, determining what research findings were suitable for classroom application, and recommending methods of dissemination. Five areas of reading were eventually examined, and an influential report was released in December 2000. This report (NRP, 2000a), *Teaching Children to Read* (Note 1), has stirred much controversy among reading experts, and both critics and supporters have been highly visible in national-level venues (e.g., Manzo, 1998; Pressley & Allington, 1999; Yatvin, 2000; Krashen, 2000, 2001; Garan, 2001, 2002; Ehri & Stahl, 2001; Shanahan, 2001; Coles, 2003). In any case, the report has played an important role in subsequent federal policy regarding reading instruction (Manzo, 2002; Manzo & Hoff, 2003).

One of the five areas of reading research examined by the NRP was phonics instruction. According to the NRP:

> An essential part of the process for beginners involves learning the alphabetic system, that is, letter-sound correspondences and spelling patterns, and learning how to apply this knowledge in their reading. Systematic phonics instruction is a way of teaching reading that stresses the acquisition of letter-sound correspondences and their use to read and spell words…. (NRP, 2000b, p. 2-89).

Using a research methodology known as *meta-analysis*, the NRP identified 38 experimental and quasi-experimental (meaning a reasonably close approximation to experimental) research studies on phonics instruction. Based on a statistical analysis of the quantitative results from these 38 studies, the NRP concluded that:

> Findings [from the meta-analysis] provided solid support for the conclusion that systematic phonics instruction makes a more significant contribution to children's growth in reading than do alternative programs providing unsystematic or no phonics instruction. (NRP, 2000b, p. 2-132)

Altogether, eleven conclusions were offered regarding the efficacy of phonics instruction, but the above finding is of prime importance.

In their deliberations on research findings, the NRP clearly recognized the ultimate need for instructional decisions to be based on the best empirical evidence and methods of analysis. At a series of five regional public hearings held prior to its work, the NRP recounted that one theme "expressed repeatedly" was

> The importance of applying the highest standards of scientific evidence to the research review process so that conclusions and determinations are based on findings obtained from experimental studies characterized by methodological rigor with demonstrated reliability, validity, replicability, and applicability. (NRP, 2000a, p. 1‑2)

Two aspects of the scientific method should be distinguished in this desideratum: the "research review process," and the "standards of scientific evidence" that led the NRP to focus on experimental studies.

In this document, we are primarily concerned with the former aspect, that is, the research review process. Most currently available reviews of the NRP's study have focused on the *interpretation* of the results for phonics instruction while assuming the basic correctness of the measurement and analytic procedures. We did not make such assumptions; rather, we designed an independent study in an attempt to reconstruct the NRP's central findings. As in other types of scientific investigation, replicability is a key criterion for judging the credibility of the NRP meta-analysis, and consequently how seriously we should consider applying its findings.

We began with the *same* 38 studies analyzed by the NRP, but in the course of our analysis, we deleted one study and added three (Note 2) others originally identified by the NRP. We then devised alternative plans for extracting and analyzing data from the 40 studies (38 – 1 + 3 = 40). Based on these analyses, conclusions were drawn and interpretations made about the efficacy of phonics instruction. Though some of the methodological steps taken by the NRP analysts were retraced, our goal was to verify whether an independent team of researchers would arrive at conclusions consistent with those in the NRP report. We did not examine how the original 38 studies were chosen. It would have been useful to examine the full range of the NRP's procedures and findings, including study selection, but this task would have required resources well beyond our means.

Our investigation resulted in several major findings. We obtained a statistically significant effect for systematic phonics instruction, but one that was substantially smaller than that of the NRP. Relative to systematic phonics, we also found that individualized instruction (i.e., tutoring v. small group or class) had a substantially larger effect while language-based instructional activities yielded a comparable effect. Finally, we concluded that there is no reason to believe that these effects are mutually exclusive. Systematic phonics instruction when combined with language activities and individual tutoring appears to have a much larger effect than phonics alone.

The remainder of this report consists of seven sections:

I. Introduction to Meta-Analysis. A brief introduction to meta-analysis is given.
II. Findings of NRP Study. An overview of the NRP findings on phonics instruction is given along with select results.
III. Reanalysis: Research Questions and Methods. Questions examined by the current study are listed, and methodological issues are described.
IV. Re-Analysis: Results. Quantitative results of the present study are given.
V. Re-analysis: Discussion. The size of the phonics effect is evaluated using results from other meta-analyses and the moderator effects estimated in the present study.
VI. Meta-analysis and Public Policy. Meta-analysis is discussed as a method for resolving controversial issues.
VII. Conclusions. Conclusions and recommendations are given with respect to integrating research, especially with respect to phonics instruction.

I. Introduction to Meta-Analysis

Meta-analysis is a *public analysis* of research findings. It uses publicly available data sources and reveals explicitly to stakeholders how data are selected and analyzed. Private knowledge of data or methodology plays no role. Cooper and Hedges (1994) summarized more elegantly:

> Two decades ago the actual mechanics of integrating research usually involved covert, intuitive processes taking place in the head of the synthesist.
> Meta-analysis made these processes public and based them on shared, statistical assumptions (however well these assumptions were met). (p. 11)

Nearly a quarter century ago, meta-analysis was developed as a set of statistical procedures for combining the results of many primary studies on a single topic (Glass, McGaw, & Smith, 1981). Previously, there was no effective way to solve the dilemmas of conflicting individual or primary studies. With meta-analysis, each study contributes information in a systematic way, and differences are resolved through statistical analysis.

In a nutshell, meta-analysis is a method of statistically summarizing *quantitative* outcomes across many research studies. Cooper and Hedges (1994) described this method as consisting of five steps:

1. *Problem formulation*. Researchers decide whether a sufficient number of studies exists for a subject of theoretical (e.g., speed of recall) or practical (e.g., class size) interest. These studies usually investigate treatments or interventions in the framework of a comparative research design. (Note 3) For example, we might ask whether students do better in a smaller class (experimental group) rather than a larger class (control group). This step also involves defining a population of interest (e.g., 4th graders) as well as measurements or outcomes (e.g., performance on multi-step math problems).
2. *Data collection: Searching the literature*. Ideally, all relevant studies would be obtained for a meta-analysis. To obtain the most exhaustive sample of studies possible, the researchers must sort through all appropriate reference systems and publications. Additional studies are frequently added by combing through the references of obtained studies as well as databases of unpublished studies. The key idea here is that if a sample of studies is obtained, that sample must fairly represent the entire population of studies to avoid bias (in the same way that the U.S. Census must ensure that hard-to-reach subpopulations are fairly represented).
3. *Data evaluation: Coding the literature*. Trained researchers must extract information about each study's results. A standard list of features (e.g., size of the treatment groups) is developed prior to reading through the studies, even though some of this information may not be reported in many studies. Different researchers who record study information work with common variable definitions so that the information is reliable and comparable across studies. (Note 4) The determination of *what counts as relevant information for coding purposes* should be made by experts who have a thorough understanding of the treatments, populations, and measurements in question. Meta-analysis requires a quantitative measure of effect or outcome, but studies using conceptually similar measures often do not use the same nominal instruments or tests. Therefore, to be able to combine quantitative treatment-control differences across instruments, they must be translated to a common scale. For example, if one wanted to add two measurements, one in centimeters and one in inches, it would be necessary to convert inches to centimeters (or vice versa). This is what an *effect size* (labeled as *d*) ideally accomplishes. It is a translation of the measured effects from different studies into comparable units (in this case, standard deviations). More description is given in Section V on the effect size measure *d*, but as a rule average effect sizes in instructional research tend to range from 0 to about ±1.
4. *Analysis and interpretation*. A central question for all comparative studies is the degree to which the experimental group (sometimes called the treatment group) outperformed the control group. Once effect sizes are computed, statistical analyses are used to estimate the average *d* and its margin of error. Analyses also determine whether certain study features like the duration of treatment influence the effect size.

Note that *estimation* of an effect is a different activity than its *interpretation*. The meaning of a measurement in centimeters can be quite different depending on, for instance, whether we are talking about the following distance of automobiles on a highway or the width of a contact lens.

5. *Public presentation*. At every stage of the meta-analysis, records should be kept regarding procedures. In reporting a meta-analysis, researchers must provide not just statistical results, but also an account of decisions that led to those results. In addition, the meta-analysis is not over until the results are linked to the research issues specified in the first step. In short, the findings must be interpreted and communicated. They must also be qualified, that is, the researchers help readers to understand limitations of the meta-analysis.

While the principles of meta-analysis are scientific, the methods it employs are not purely formulaic. Human judgment is a key element in each of the five steps. In particular, meta-analysts rely on expert judgment for converting narrative descriptions of a study's treatments and subject populations to quantitative measurements. Such coding often requires substantive expertise in addition to research and quantitative skills. (Note 5)

II. Findings of NRP Study

The subgroup of the NRP for Phonics Instruction described the five steps of its meta-analysis in Chapter 3, Part II of *Teaching Children to Read*. In particular, 11 major conclusions were listed (NRP, p. 2-132 to 2-136). The report is well-summarized by Ehri et al. (2001, abstract):

> A quantitative meta-analysis evaluating the effects of systematic phonics instruction compared to unsystematic or no phonics instruction on learning to read was conducted using 66 treatment-control comparisons derived from 38 experiments. The overall effect of phonics instruction on reading was moderate, $d = 0.41$. Effects persisted after instruction ended. Effects were larger when phonics instruction began early ($d = 0.55$) than after first grade ($d = 0.27$). Phonics benefited decoding, word reading, text comprehension, and spelling in many readers. Phonics helped low and middle SES readers, younger students at risk for reading disability (RD), and older students with RD, but it did not help low achieving readers that included students with cognitive limitations. Synthetic phonics and larger-unit systematic phonics programs produced a similar advantage in reading. Delivering instruction to small groups and classes was not less effective than tutoring. Systematic phonics instruction helped children learn to read better than all forms of control group instruction, including whole language. In sum, systematic phonics instruction proved effective and should be implemented as part of literacy programs to teach beginning reading as well as to prevent and remediate reading difficulties.

For additional detail with regard to the overall results, we give the complete text of the first conclusion from the NRP report:

> Children's reading was measured at the end of training if it lasted less than a year or at the end of the first school year of instruction. The mean overall effect size produced by phonics instruction was significant and moderate in size ($d = 0.44$). Findings provided solid support for the conclusion that systematic phonics instruction makes a more significant contribution to children's growth

> in reading than do alternative programs providing unsystematic or no phonics instruction. (NRP, 2000b, p. 2-132).

Data analyses supporting these conclusions were based on a straightforward design: treatment groups receiving systematic phonics were compared to control groups receiving unsystematic or no phonics instruction. Yet both the experimental and control groups might receive *mixtures* of phonics, language instruction, and other activities. The NRP did examine whether the effect of phonics instruction was influenced by moderator variables, such as socio-economic status or phonics programs. However, no attempt was made to classify the degree of phonics or the mixtures of phonics and other language activities in the groups being studied.

Treatment and Control Group Definitions

In order to understand the overall effect (d = .41/.44), it is necessary to understand the characteristics of the treatment and control groups (Note 6). The NRP described treatment groups as including systematic phonics instruction while control groups, though they may have had some phonics instruction, as having various other types of instruction (NRP, 2000b, p. 2-103) with *less* systematic phonics. Thus, the effect size generally signifies the advantage of *more versus less* systematic phonics instruction:

> Whereas some groups were true “no-phonics” controls, other groups received some phonics instruction. It may be that, instead of examining the difference between phonics instruction and no phonics instruction, a substantial number of studies actually compared more systematic phonics instruction to less phonics instruction. (NRP, 2000b, p. 2-124)

Because almost all children received some instruction in phonics during the course of comparative studies, this formulation is realistic. However, the degree of phonics instruction varied from study to study, and it is possible that a treatment in one study could resemble a control in another.

While we believe that the effect size can be a useful measure in such situations, it must be realized that any ambiguity in how comparisons vary across studies adds some ambiguity to the interpretation of the overall or average effect size. The NRP surmised that the effect of such treatment-control variability might be to *underestimate* effect sizes. In many cases, however, children receiving systematic phonics instruction were also receiving activities consistent with the aims and purposes of whole language. Thus, uncontrolled mixtures might also serve to *overestimate* the effects of phonics instruction.

Others have written about the false dichotomy between language and phonics instruction (e.g., Fletcher and Lyon, 1998). (Note 7) A number of phonics instruction treatments are described in the NRP report including synthetic, analytic, analogy, onset-rime, phonics through spelling (NRP, 2000b, p. 2-99), and embedded phonics. Many contain some degree of language instruction. For example, although “embedded” phonics was not defined in the NRP report, Foorman, Francis, Fletcher, Schatschneider, and Mehta (1998) described their “embedded code” treatment as including “whole-class activities such as shared writing, shared reading, choral or echo reading, and guided reading” (p. 40). In addition the teachers would “frame a word containing the target spelling pattern during a literacy activity” (p. 40). Consequently, the treatment is consistent in some important respects with language-based instruction, though it can also be described as a type of phonics instruction. While such treatments defy simple labels, they can be coded on various dimensions that more accurately describe the "package" of treatment conditions. Analyses can then be undertaken to sort out the unique effects of various instructional activities and conditions.

Outcome Variables and Units of Analysis

The NRP subgroup on phonics instruction computed effects sizes for dependent variables that fit into one of 7 categories (also see Table 1) (Note 8):

1. Word ID
2. Decoding
3. Spelling
4. Comprehension
5. Nonword reading
6. Oral reading
7. General reading

Table 1
Dependent Variable Categories.

Category	Label	NRP Label
1	decoding regular words	decoding
2	decoding nonwords	nonwords
3	sight word ID	word ID
4	spelling	spelling
5	comprehension	comprehension
6	oral reading	oral reading
7	general reading	general reading
8	language	*
9	phonemic awareness	*
10	alphabetic knowledge	*
11	vocabulary	*
12	writing	*
*Category not used in NRP study		

For each category within each treatment-control comparison, it is our understanding (NRP, 2000a, p. 1-10) that either mean or median effect sizes were computed for each cohort of students when results for more than one test instrument were available. In some cases, studies did not report measures for some categories, in which case the category was left blank (i.e., a “missing value”) in Appendix G. At most, one effect size was reported for each category for each cohort/comparison.

Importantly, measures were excluded from this classification if they were used during (or as part of) phonics instruction (NRP, 2000b, p. 2-110). Such effect sizes would be expected to be larger due to “teaching to the test.” No distinction was made between standardized and experimenter-devised tests. Because standardized tests are targeted to a wider range of ability, the NRP surmised that they might be less sensitive to change and thus “underestimate effect sizes slightly” (NRP, 2000b, p. 2-111).

Criticisms of the NRP Meta-Analysis

Three prominent criticisms of the NRP meta-analysis of phonics instruction have spurred public debate. The first concerns methodology; the second concerns the link between evidence and conclusions; and the third, the procedures with which research activities were conducted.

The first criticism is that a narrow population of children was represented in the 38 studies that comprised the meta-analysis (Garan, 2002). In particular, Garan argued that many of the studies did not include “normal readers” and none included groups of advanced readers. Thus, it would be difficult to generalize the findings broadly across typical populations of students. The second criticism is that the term “reading” was not used in a consistent manner; the term reading can refer to simple “word calling” (e.g., a response to the question “Can you say this word?”), but it can also refer to the ability to derive meaning from connected text (Yatvin, 2002). If it is said that “Phonics instruction improves reading,” it is important to know what kind of reading is signified. The third criticism was that the process used to conduct and report the meta-analysis was flawed. According to Yatvin (2002), the NRP study on phonics instruction was completed in a very short time. In October, 1999, five months before the due date, a determination was made that the completion of the study required resources beyond the capacity of panel members, and it appears that a researcher who was not a member of the NRP was commissioned to conduct the meta-analysis. (Note 9) Upon completion of the study, again due to time constraints, the panel originally in charge of designing and conceptualizing the research had only four days to review the final report before it went to press. Yatvin also observed that only one panel member (Yatvin) had teaching experience, and thus the NRP had little expertise for the purpose of linking research findings to practice.

The NRP addressed some of these issues. The 38 studies provided 66 (Note 10) treatment-control comparisons, and of these, 23 comparisons included normal readers (about 35%). In regard to the second criticism, the NRP found that:

> The majority (76%) of the effect sizes involved reading or spelling single words while 24% involved reading text. The imbalance favoring single words is not surprising given that the focus of phonics instruction is on improving children’s ability to read and spell words. (NRP, 2000b, p. 2-92)

Even from this brief quote, it is clear that a necessary distinction must be made between “word reading” and conceptualizations of reading that imply understanding of connected text. “Word reading” is just one connotation of reading, yet the distinction isn’t maintained consistently in formal documents. For example, in Ehri and Stahl’s (2001) rebuttal to Garan

(2001) (Note 11) they reported that clear evidence was found to support the conclusion that

> Systematic phonics instruction was found to be more effective than unsystematic phonics instruction or no phonics instruction in helping students learn to *read* [emphasis added]. (Ehri and Stahl, 2001, p. 18)

One could define reading as "reads single words in isolation," which would be consistent with the NRP's data analyses. But reading could also be defined as "reads connected text," that is, sentences or stories. Obviously, one's sense of the study's outcome—as represented in the above quote—depends almost entirely on how reading is defined.

The third criticism was that not enough time was allotted to carry out the charge of Congress, and that the final report was not subjected to formal review. In fact, the study was under intense time pressure from inception. According to Yatvin, who wrote a minority addendum to the final report,

> In fairness to the Panel, it must be recognized that the charge from Congress was too demanding to be accomplished by a small body of unpaid volunteers, working part time, without staff support, over a period of a year and a half. (The time Congress originally allotted was only 6 months.) (Yatvin, 2000, p. 2)

Whether the resources and time were sufficient to carry out such an important study is now a moot issue. The question of interest is whether the meta-analysis conducted by the NRP is sufficiently reliable and valid for guiding instructional policy in early reading. In the present study we address the topic of whether the central NRP results can be replicated by a different team of analysts. A successful replication would provide convincing evidence of accuracy and allay concerns about study logistics.

III. Reanalysis: Research Questions and Methods

The NRP results were given for 11 central questions regarding phonics instruction. In this re-analysis, we will be concerned primarily with two of these: "Does systematic phonics instruction help children to learn to read more effectively than nonsystematic phonics instruction or instruction teaching no phonics?" (NRP, 2000b, p. 2-132); and "Is phonics instruction more effective when it is introduced to students not yet reading, in kindergarten or 1st grade, than when it is introduced in grades above 1st after students have already begun to read?" (NRP, 2000b, p. 2-133).

Using public—that is, published—accounts of data and methodology, we re-examined the evidence offered by the NRP on the efficacy of phonics instruction. We designed an effect size database, recomputed effect sizes for all outcomes available, and then carried out analyses in which effect sizes were related to study characteristics. One study by Vickery, Reynolds, and Cochran (1987), which is described in Appendix C, examined the effect of the *same* treatment on remedial and nonremedial students. Because there was no control group, we deleted this study from our database (see inclusion criteria on p. 2-108 to 109 in NRP, 2000b). We included another three studies that were identified by the NRP but not included in their meta-analysis. These are described in Appendix A of this report. Thus, our database was constructed from 40 studies originally identified by the NRP; however, the merits of the original NRP sample or sample selection process is beyond the scope of the present study. (Note 12; Note 13)

Our analytic strategy had several components. We selected a unit of analysis, defined alternative weighting schemes, and used multiple regression to identify the unique contributions of variables that moderate the treatment. By *moderator* variable, we mean a component of treatment delivery that leads to a stronger or weaker effect. Four new moderator variables were constructed for specifying the treatment conditions: the degree of phonics systematicity; degree of coordinated language activities; whether treatments were regular in-class or pullout programs; and whether basal readers were used. These variables, which were coded from the research studies by means of rubrics, provided the explanatory power missing from the simple comparative design used in the NRP analyses. That is, the NRP design did not fully account for variation in the mixtures and degrees of treatment delivered to both experimental and control groups. Other moderators were borrowed from Appendix G of the NRP report. Using regression analysis, we then predicted treatment outcomes (i.e., effect sizes) with the four new moderators and: the size of the instructional unit (tutoring, small groups, class); whether treatment conditions were randomly assigned; whether standardized tests were used; and the age (Note 14) of students.

There are two important design facets in a meta-analysis. The first is a design for *data collection*, while the second parallels the usual sense of the word in the phrase *experimental design*. That is, there is one design for data collection, and another for analysis. In order to address the weaknesses of the simple comparative design of the NRP study, we coded moderator variables, but we also planned for a more complete use of the information within each of the 40 studies. In particular, we distinguished untreated control groups from "alternative" treatments, and included both, as described below. This can be likened to filling out the cells of—or balancing—an experimental design, while the increasing the number of studies adds to sample size. The recognition of this distinction is not evident in the NRP analytic plan.

Database Design

Including Groups for Comparison. As noted above, in each study the NRP designated as the control a group with less systematic phonics than the treatment group (or groups). Ironically, this procedure in some cases led to ignoring information from groups labeled as "control" by the authors of the primary studies. For example, in the study by Lovett, Ransby, Hardwick, Johns, and Donaldson (1989) three groups were used: the Decoding Skills Program (DS), the Oral and Written Language Stimulation program (OWLS), and a Classroom Survival Skills program (CSS). The third group was described as a "control procedure in which subjects received the same amount of clinic time and professional attention as those in the experimental remedial programs" (p. 96); however, CSS students received training in activities that didn't include reading. It appears that non-treatment controls such as the CSS group were excluded from the NRP study when programmatic controls like OWLS were present. Thus, the NRP effect sizes for Lovett et al. (1989) are based solely on the comparison of the DS to the OWLS program.

In such cases, we computed effect sizes for DS versus CSS *and* OWLS versus CSS. However, we coded (with treatment indicators determined by rubric codings) the DS program as having systematic phonics instruction while the OWLS program was coded as language-based. This strategy yields an important source of information for disentangling treatment effects because untreated control groups can provide a common basis for comparison across studies. The component effects of treatment mixtures may then be more accurately identified.

Defining Control Groups. More than one control group may have been available for computing effect size. For example, in Foorman et al. (1998), there were four groups described as: direct code (DC), embedded code (EC), implicit code-research (IC-R), and implicit code-standard (IC-S). It appears that the NRP analysts used the IC-S group as the control even though the authors of the study asserted that comparisons among IC-R, DC, and EC provided the most relevant information about instructional differences because the IC-R group controlled for teacher training.

We decided to use the IC-R group for computing effect sizes based on the general rationale in this paragraph, which we used for all studies. The most valid control was taken as the group that received the same kinds of treatment activities (e.g., individual attention, duration of treatment), but not the treatment itself—either language or phonics. This would serve to control for as many background variables and moderators as possible. For instance, if there were a choice of control between two groups that did not involve phonics or language instruction, then we would use this rule to choose the control. We coded systematic language programs as treatments unless there was not another control group available. In a study with only a phonics group and a language group, we compared the phonics to the language group to obtain the effect size, but coded the comparison as being Phonics v. Language rather than Phonics v. Control. At least three possible classes of comparison (phonics-control, language-control, and phonics-language) were defined by the rubric indicators.

In summary, we included control groups having no systematic phonics or language interventions, whereas the NRP analysts did not. However, when two control groups were available, we chose the one most like the treatment group in terms of characteristics ancillary to the intervention.

Coding Rubrics and Inter-Rater Reliability

We coded the characteristics of both treatment and control groups with rubric indicators. The rationale for this practice is that coding is a measurement process, requiring inference, and not a simple reading of a study. Since coding is a measurement process, its scientific warrant should be established by demonstrating inter-rater agreement. The credibility of the limited moderators coded by the NRP team was also established by demonstrating high inter-rater agreement.

In Table 2, the rubrics are given that were used to code treatment characteristics. We distinguished among three levels of phonics instruction; two levels of language; basal reader usage; and supplemental/pullout versus regular in-class instruction. Rubric codings provide a richer quantitative description of studies in which instruction is comprised of mixtures of phonics, language, and other elements. For each study, three independent codings were obtained. The first codings were given by the authors of the present study, each of whom had participated in all aspects of at least one previous meta-analysis. None had previously participated in a study of phonics or whole language instruction, and none had taken a public position in the phonics versus whole language debates. The second and third codings were provided, respectively, by an experienced reading teacher and a university professor, each with a national reputation in reading instruction.

Table 2
Rubrics for Coding Treatment Conditions.

Phonics	
ng	No information in study to infer code.
0	No specific phonics intervention was given. In most cases, we know that it is highly probable that students received some kind of phonics activity, especially for longer interventions. Moreover, even if no phonics instruction was associated with the treatment delivered, it may have been the case that other instructional activities (external to the treatment) included phonics. In short, we were not able to distinguish among these possibilities.
1	Treatment specifically included phonics activities, but treatment activities were not described in detail as being direct, systematic instruction. Organized phonics were embedded in language instruction.
2	Treatment was described as including direct, systematic phonics instruction. It was most often the case that this description specifically included blending.
Replace	
ng	No information in study to infer code.
0	Treatment did not replace regular classroom instruction. In some cases, the treatment consisted of a supplemental program. For example, students received treatment at facilities outside of schools (e.g., hospital setting on Saturdays).
1	Treatment was regular classroom instruction, or the treatment completely replaced regular classroom instruction.
Basal	
ng	No information in study to infer code.
0	Basal reader was not used.
1	Treatment was described as including a basal reader, or it was highly probable that a basal reader was used. For example, a 4-year treatment consisting of regular classroom instruction almost certainly used a basal reader at some point, even if it was not specifically mentioned.
Language	
ng	No information in study to infer code.
0	No systematic or formal language activities were included.
1	Language-based (non-basal) treatment was given. This may have consisted of whole word or whole language programs.

For each effect size computation, both experimental and control groups were coded according to the rubrics, allowing for the possibility that any group could be coded as having both phonics and language instruction. However, no phonics treatment labeled as such ever had less systematic phonics instruction than the group chosen as the control, though both groups may have had language instruction. In some cases, study information for coding a rubric was denoted as “not given” by one or more coders. Our guiding principle on this matter was that evidence of “presence” was required in order to make inferences regarding the effects of a rubric variable. We converted “not given” responses to zeros. For example, if a study did not report that basal readers were used, but it was known that the reading program formally included basal readers (and did during the timeframe of the study), then assuming their presence was a relatively safe inference. However, for less familiar or unknown reading programs, it was safest to assume basal readers were not used. In short, the conservative approach to coding was to require evidence of “presence” rather than “absence” when linking treatment or moderator indicators to study outcomes.

In Tables 3a-3d, agreement analyses are given for each of the four rubric variables separately. Under the column labeled “Judges Codings” the number of each possible combination (i.e., unordered triplet) of three codes, one for each judge, is given. Overall, there was substantial agreement among coders, given the evidence-of-presence requirement. In addition to the data in Tables 3a-3d, it is also useful to consider that three raters operating at random with 95 total comparisons would only have an expected value of about 10-11 matches with a 3-point rubric, and only about 23-24 on a 2-point rubric.

Table 3a
Inter-rater Agreement for the Phonics Rubric
(Cronbach’s alpha for this rubric was .95)

.

Judges Codings	***n* (95 total)**	**Cumulative Percent**	**Agreement Type**
0,0,0	22	23	Perfect
1,1,1	13	37	Perfect
2,2,2	31	69	Perfect
0,0,1	12	82	Adjacent
0,1,1	6	88	Adjacent
1,1,2	4	93	Adjacent
1,2,2	5	98	Adjacent
0,1,2	1	99	———
unclassed	1	100	———

Table 3b
Inter-rater Agreement for the Language Rubric
(Cronbach’s alpha for this rubric was .79.)

Codings	***n* (95 total)**	**Cumulative Percent**	**Agreement Type**

0,0,0	55	58	Perfect
1,1,1	11	69	Perfect
0,0,1	9	79	———
0,1,1	19	99	———
unclassed	1	100	———

Table 3c
Inter-rater Agreement for the Basal Reader Rubric
(Cronbach’s alpha for this rubric was .82.)

Codings	***n* (95 total)**	**Cumulative Percent**	**Agreement Type**
0,0,0	52	55	Perfect
1,1,1	18	74	Perfect
0,0,1	19	94	———
0,1,1	5	99	———
unclassed	1	100	———

Table 3d
Inter-rater Agreement for the Pullout Rubric
(Cronbach’s alpha for this rubric was .87.)

Codings	***n* (94 total)**	**Cumulative Percent**	**Agreement Type**
0,0,0	30	32	Perfect
1,1,1	40	74	Perfect
0,0,1	12	87	———
0,1,1	11	99	———
unclassed	1	100	———

When we encountered a difference among coders, the final code was chosen as the consensus code in almost all cases. In the few cases where a 2-of-3 majority was not obtained, codes were averaged. For example, in one comparison the level of phonics was given codes of 0, 1, and 2, and in this case, the results were averaged resulting in a code of 1.0. Given the “majority rules” principle (Orwin, 1994), the three judges were overruled 24, 25, and 56 times out of 404 coding instances. This translates into overruled percentages of about 6%, 6%, and 14%, respectively. Thus, an individual judge’s code was retained in a minimum of 86% of the coding instances. The coefficient alphas were relatively high at .95 (degree of phonics), .79 (presence of language activities), .82 (use of basal readers), and .87 (regular v. pullout program).

The effort in detailing treatment conditions is important for making a strong statistical link between treatments and outcomes, and would ideally be planned in the design of the study and coding protocol. Even so, a dose of realism is required in this effort. The typical study examined lacked clarity with regard to treatment conditions. It was not uncommon that a total of three or four sentences were devoted to describing an intervention. Though we feel 69% is an acceptable rate of *perfect* agreement for level of phonics instruction (with a reliability of α = .95, see Table 3a), at least part of the 31% of disagreement can be attributed to the lack of clear descriptions of independent variables. Some "measurement error" reflects ambiguous descriptions rather than ambiguity inherent in the judgment process.

In the analyses reported below we used, but did not code, other moderator variables in addition to the rubric indicators. These moderators were borrowed directly from the NRP study including treatment unit, age/grade, SES, and reading ability. We coded, but did not obtain inter-rater agreements, for several additional variables for each effect size in our database including the size (*n*) of each treatment/comparison group, whether the effect size was from a randomized study, and whether the effect size was from a standardized instrument.

Dependent Variables

Variable Categories. All effect sizes were recomputed for all available outcome measures that could be considered as falling into one of the categories in Table 1. In some cases, we considered outcomes that the NRP did not use, such as alphabetic knowledge, which refers to how well students can connect phonemes to graphemes. Though these measures fell outside the range of the NRP's definition of reading, we felt the information was useful.

Effect Size Computation. Most criticisms and counter-criticisms of the NRP report accept as their starting point the computed effect sizes as obtained by the NRP analysts. We did not use the published NRP effect sizes because independent computation is more consistent with the goals of a validation study based on the merits of replication. Therefore, one major focus of the present study is computational: Can the general effect size obtained by the NRP analysts be replicated? However, since we recomputed effect sizes based on a different design (than the one used by the NRP) for experimental-control comparisons, a one-to-one comparison was not possible. For this purpose, we devised an approximate method of comparison (described in Section IV).

Computing an effect size can pose a difficulty with which meta-analysts are all too familiar, but one that may not be transparent to a consumer of meta-analytic information. In studies that do not report the necessary information for a simple computation, information must be pieced together—sometimes using specially designed procedures that may require a number of assumptions. In this section we review a number of these issues that are pertinent to the studies on phonics instruction. Although the DSTAT program was used for computations (Johnson, 1989) by the NRP team, it is often the case that judgments must be made as to what information to enter into the program; different choices may yield different results even when calculations are error-free. In a number of instances, the NRP team may not have appreciated the complexities of computing the effect size *d*, or they did not provide rationales for their methods. In this regard, we provide several clarifications below for facilitating accurate effect size computation.

Although the NRP cites Cooper and Hedges (1994) regarding formulas for computing effect sizes, the basic formula given by the NRP (NRP, 2000a, p. 1-10) and reproduced below is incorrect:

$$(1) \quad \frac{\overline{X}_T - \overline{X}_C}{.5(s_T + s_C)}$$

Compared to the pooled effect size estimator (*g*) given by Hedges (1985, p. 78)

$$(2) \quad \frac{\overline{X}_T - \overline{X}_C}{\sqrt{\frac{v_T s_T^2 + v_C s_C^2}{v_T + v_C}}}$$

we can see that while the numerators of (1) and (2) are the same, the denominators differ (v_T and v_C are the degrees of freedom for the experimental and control groups, respectively). Moreover, it can be shown that the effect size given by (1) is always larger than that given by the standard formula in (2). The magnitude of this difference is not large, however, and the NRP calculations appear to have been performed with the correct formula. Nevertheless, it is important to communicate established procedures in a public document. If a nonstandard formula is used, a justification should appear in text, but we know of no justification for the formula in (1). (Note 15)

For the most part, we computed effect sizes in a manner consistent with general methodological descriptions given in the NRP; the Hedges correction was used in all cases (Hedges & Olkin, 1985, p. 81). However, because few in-depth details were provided (e.g., NRP 2000b, pp. 2-110 to 2-111), we used several additional guidelines for the current study:

> a) Standard deviations were pooled across all posttest treatment and control groups within a cohort of students to create a common denominator. Hedges effect size adjustment was applied to *g* to arrive at *d* (using degrees of freedom based on the pooled sample).
>
> b) When pretest means were available, effect size numerators were computed as differential average gains to help control for pre-existing differences. Effect sizes, according to the first guideline, were then obtained via division with a common posttest standard deviation. If covariance adjusted effects were reported, these were used in the numerator instead of the difference between average gains of treatment and control groups.
>
> c) When testing was carried out on more than two occasions during a treatment intervention, we computed gains based on pretest and immediate posttest means. If a treatment spanned several years (or grades), we computed an effect size for the first year using the second guideline. For each ensuing year separately, we computed an effect size using the previous year's posttest as the following year's pretest.
>
> d) Effect sizes were computed with custom programming developed for each individual study rather than using one of the available software products. For the most part, calculations were based on formulae given by Cooper and Hedges (1994).

Units of Analysis. In some cases, classrooms or even schools are used as the units of analysis rather than individual students, and this phenomenon did occur in the set of phonics instruction studies. In this case, classes (or schools) are the *units* of observation, and class means comprise the data to be analyzed. The formula given in (2) typically pools individual level standard deviations that are first calculated with the formula:

(3) $$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

When individual observations are group means, however, the estimate of variability, *s'*, is

(4) $$s' = \frac{s}{\sqrt{n}}$$

which is the formula for the standard error of the mean. Upon comparison, it can be seen that (4) will be smaller than (3) depending on *n*, which is the class (or school) size. Therefore, an effect size using class means will be larger than one based on individual student scores by the multiplicative factor √n. With moderately sized classes, the use of means can result in substantially larger effect sizes, but *these are not comparable* to the effect sizes of other studies whose units of observation are students. To remedy this disparity, effect sizes must be translated to the individual metric. (Note 16) As we shall see below, the greatest discrepancy (between a recomputed and original NRP effect size) was due to a unit of analysis problem.

Weighting Studies

In the NRP study, effect sizes were computed for each experimental treatment. For example, if there were one outcome variable, two distinct phonics-based treatments (A and B), and one control group (C), two effect sizes would be computed (A versus C, and B versus C). If there were two or more outcome variables in a dependent variable category (e.g., two spelling tests), the effect sizes for A-C and B-C would be averaged separately within this category.

Because the NRP reported 66 comparisons from 38 studies, some studies contributed more than one effect size. For example, one study by Vickery et al., 1987, contributed 8 comparisons—4 grade level cohorts crossed with two levels of remediation. There are a number of methods for computing the overall average *d* in this situation. First, one could compute the simple average across the 66 comparisons given in Appendix G of the NRP report (NRP, 2000b, pp. 2-169 to 2-175), which results in a mean of .46. This is close to the value .41 which was reported by Ehri, Nunes, Stahl, and Willows (2001). Implicit in this procedure is that the Vickery et al. (1987) study receives 8 times the weight of a study that contributed a single effect size—because it examined 8 distinct treatment-control cohorts. (Note 17)

A second method consists of weighting studies by the total *n* of the comparison (treatment + control); in other words, comparisons with larger *n*s would receive more weight. This was the method used by the NRP, and results in mean *d* = .41. In the NRP study, the rationale was given that

> The subgroups [committees] weighted effect sizes by numbers of subjects in the study of comparison to prevent small studies from overwhelming the effects evident in large studies. (NRP, 2000a, p. 1-10).

With this practice, however, large studies overwhelm small studies. For example, in Gersten, Darch and Gleason (1988), data from 1973-1974 are available for two cohorts of children with a total $n = 242$. Treatments were provided at the class level. In contrast, one comparison described by Gillon and Dodd (1997) contains $n = 10$ students in two groups. Given a simple weighting by n, the latter study would have about 1/24th the weight of the former study. It is our opinion, that this weighting practice should not be automatic; application of statistical weights necessarily gives studies using classes more weight than studies using small groups or tutoring.

In a third method for averaging across studies, separate *studies* are given equal weight. In this case, the 8 effect sizes in Vickery et al. (1987) would each receive a weight of 1/8; and the weights would sum to 1.0. This weighting practice would be repeated for all studies resulting in a set of weights that would sum to exactly the number of independent studies. In the NRP study, this weighting procedure results in an mean $d = .54$. (We note that this effect is *larger* than the estimate reported in *Teaching Children to Read*, but stay tuned.) In our opinion, this approach makes sense when a set of effect sizes is relatively homogenous. Though Shadish and Haddock (1994) asserted that "all things being equal," weighting sample size is the most widely accepted practice, Hedges and Olkin (1985) cautioned that statistical weights should be considered only in cases with homogenous effects sizes:

> Before pooling estimates of effect size for a series of *k* studies, it is important to determine whether the studies can reasonably be described as sharing a common effect size. (p. 122)

Thus, "all things being equal" can be accurately interpreted as "sharing a common effect size."

A fourth method of weighting represents a compromise between statistically weighting and equally weighting studies. Let the statistical weights be labeled as *WGT1*, and let the equal representation weights be labeled *WGT2*. A compromise between the two weight types can be achieved by taking *WGT3 = WGT1*WGT2*. In the latter approach, consideration is given both to study representation and sample size. See Table 4 for definitions of the three types of weighting. For the analyses in the present report, we examine regression estimates derived from the weighting systems represented by *WGT1* and *WGT3*.

Table 4
Definitions of Alternative Unit Weights:
Equal Representation, Optimum, and Compromise.

WGT1:	If a single study contributed *k* records to the aggregated database, the *equal representation* weight was defined as: $WGT1 = \frac{1}{k}$

WGT2:	For a particular record in the aggregate database, the total number of observations for the treatment and control groups was: $n_{TOT} = n_T + n_C$ The weight was then taken as n_{TOT}. Rather than using this approach, we opted to use the *optimum* weight defined by Hedges and Olkin (1985, pp. 86 & 110) as: $WGT2 = \left(\frac{n_{TOT}}{n_T n_C} + \frac{d^2}{2(n_{TOT} - 2)} \right)^{-1}$ After examining the distribution of *WGT2*, we set a maximum value so that the highest values were no more than 15 times larger than the smallest values.
WGT3:	Given *WGT1* and *WGT2*, this *compromise* weight *WGT3* was computed as: $WGT3 = WGT1 * WGT2$

How studies should be weighted is a critically important issue, because different weighting methods may give different results. Ironically, the NRP choice resulted in large studies effects overwhelming those of smaller studies, and the consequences of using this choice should be carefully considered. In our database of effect sizes, the test of homogeneity was highly significant (Q = 813.46, 223 df; equivalent z-statistic is roughly z = 27.94) indicating that the studies did *not* share a common effect size (i.e., the hypothesis of homogeneity was rejected). Statistical weights may be inappropriate for the NRP data because of potential qualitative differences between small and large studies. Moreover, some studies included multiple comparisons, and statistical weighting gives such studies many times the influence of studies with a single comparison. A procedure in which studies are given equal weight may provide the most "equitable" reading of the experimental literature, but a compromise, which balances representation and statistical precision, may also be useful. Other procedures may also be defensible, but in any case an explicit justification should be provided.

The issue of weighting involves notions that are fundamental to the ideals of meta-analysis. Though it can be understood as a statistical issue, weighting can also be understood relative to the questions "What counts as evidence?" and "How should evidence be accumulated?" What counts as research in education usually comes in the form of a "study" in which an author analyzes data and reports conclusions based on those analyses. The results from a single study may be extremely trustworthy and valuable, and so a problem arises when we wish to "sum" the evidence in two or more studies. Other things being equal, conclusions from two studies using the same data would not be valued equally to conclusions from two independent studies. Yet the problem is not simply to determine appropriate weights for different studies, but how to understand the role of cumulative evidence vis-à-vis the role of in-depth knowledge flowing from a single, well-executed study.

Meta-analysis is a systematic method for summarizing the knowledge inherent in a research literature. Information concerning study outcomes based on unreported or private knowledge can obviously not add to this summary, even though what is actually learned from a study *does* include unreported and private knowledge. The truth of analytic conclusions can only be linked to "reports" of empirical investigations. This assertion is very different from the "garbage in—garbage out" axiom, which implies that a simple "truth in—truth out" model is possible for synthesizing research studies. The process of establishing warrants for conclusions is different in meta-analysis than it is in primary research since in published primary studies authors have direct access to contextual information (e.g., vested interests) that is not printed, but nonetheless influences reported conclusions. One important assumption of meta-analysis is that the effects of unreported information will "average out" across independent studies. This is why fair representation and appropriate weighting strategies are such important prerequisites to valid conclusions.

Case Studies

The quality of any meta-analysis is fundamentally based on studies that meet inclusion criteria. In the NRP phonics instruction meta-analysis, the foremost criterion was that "Studies had to adopt an experimental or quasi-experimental design with a control group." (NRP, 2000a, p. 2-108). In addition studies had to appear in a refereed journal after 1970, had to provide information for testing the efficacy of phonics instruction on reading, and had to report statistics necessary for computing effect sizes. Having obtained such studies, information was coded and analyses were conducted. The goal of the NRP meta-analysis was to identify reliable and replicable results in the area of early reading.

In Appendix B, we provide a perspective on three studies that met these inclusion criteria. Our goal is to provide readers with a deeper familiarity with the literature, one that extends beyond the typical boundaries of a meta-analysis. This is important for illustrating how well the inclusion criteria performed in obtaining methodologically rigorous studies, and for giving a more salient notion of the confidence with which we can generalize. Indeed, cases studies were also included in the NRP report because of their descriptive value. We emphasize that the studies in Appendix B of the present report are given for the purpose of illustration—issues arise with any study put under a microscope.

The case studies serve to illustrate methodological issues in a number of areas including: choice of control group, unit of analysis, and study selection criteria. While all three studies use quasi-experimental designs, a more in-depth examination of these can facilitate a practical understanding of the variety and limitations in this design approach to reading research. In our judgment, these studies are representative of, if not of higher quality than, the entire set of 40 studies.

IV. Re-Analysis: Results

The NRP used cohort comparisons as the unit of analysis, and then applied statistical weights. (Note 18) We used two strategies for weighting. (Note 19) The first strategy is the meta-analytic equivalent of "one person one vote" representation. In the second compromise strategy, we combined statistical (inverse variance or comparison *n*) weights with equal representation weights. We remind the reader that the usefulness of weighting—as well as that of the entire meta-analytic enterprise—depends on how well a set of studies represents the research literature.

The process of coding resulted in obtaining 491 effect sizes from 40 studies for 12 dependent variable (DV) categories. The Vickery study (described in Appendix C) was deleted due to lack of a control group. This left 37 original NRP studies, to which we added three studies with phonemic awareness outcomes. It appeared to us that the latter three studies (described in Appendix A), which were identified but not included in the NRP analysis, met the inclusion criteria of the NRP. Each of these studies, which contributed 7 records total to the database, included at least one reading outcome from categories 1-6 in Table 1. This database did not include effect sizes from follow-up comparisons. The data file contained one data record for each effect size. However, single studies often contributed more than one effect size for a DV category. Moreover, a *d* for the same outcome variable category might be computed for more than one cohort within a study. To manage this redundancy, we aggregated effect sizes to the comparison level within each DV category within a study. This resulted in a primary analysis file of 225 observations (out of a possible 480, which equals 40 studies multiplied by 12 DV categories); 60 of these represented DV categories not included in the NRP study. For each case in the aggregated file we included moderator variables such as duration of treatment, size of the treatment unit, rubric codes, and the like.

Our unit of analysis was "comparison." That is, if a study compared one treatment to one control group, and measured two outcomes, then there were 2 effect size records for one comparison. In some cases, a study had two treatment groups (T_1 and T_2) and one control group (C); in this case with two outcomes, there were 4 effect size records (T_1 v. C, and T_2 v. C. crossed with two outcomes). Equal representation weights were then obtained as the inverse of the number of records per study. Multiple cohorts were averaged, if they existed, within comparison unless the treatment conditions changed across time.

Agreement with the NRP Study

Because of the design difference between the NRP meta-analysis and the present reanalysis, it is not possible to compare the effect sizes for the two studies directly. However, if effect sizes are aggregated to the study level (excluding studies with TP = 0), we can examine the consistency of the two sets of effect sizes. In Figure 1, the scatter plot shows that two studies (labeled 12 and 53) appear as outliers. Study 53 contained an (d = 8.79) outlier and appeared to have been removed from most, if not all, NRP calculations. In study 12, effect sizes were computed by the NRP team with class means; the required conversion of the pooled standard deviation to the individual metric was not made. With these two studies removed, r^2 = .754 for effect sizes based on the original 7 NRP categories (e.g., Nonwords, Decoding, etc.) with TP=1 or TP=2.

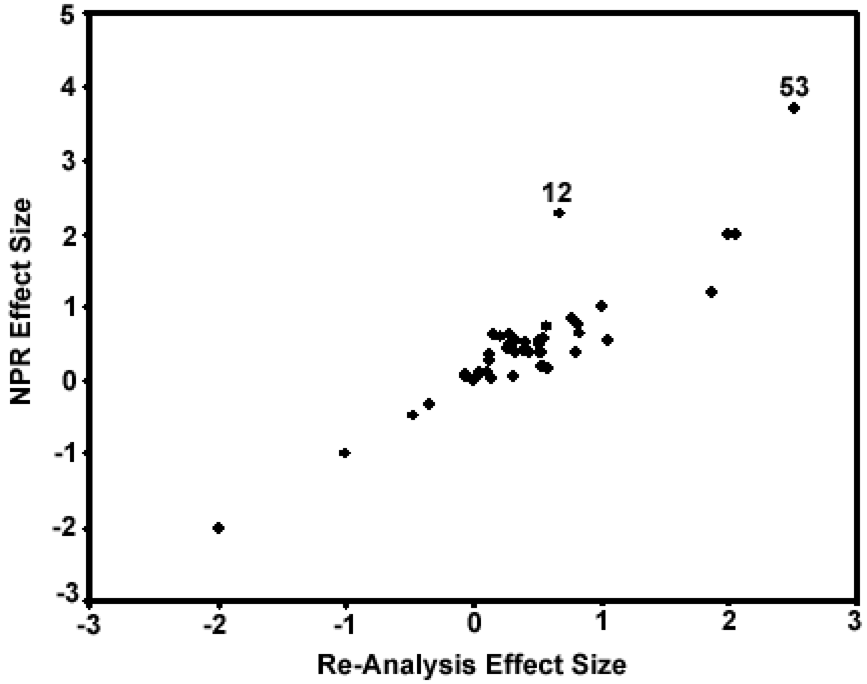


Figure 1. Scatterplot of NRP Calculated Effect Sizes and Our Re-Analysis Calculated Effect Sizes

Again with studies 12 and 53 removed (as well as Vickery et al., 1987), the overall averages of these two sets of effect sizes do not significantly differ using a paired samples t-test (t = -.447, p = .658, 33 df). Clearly, the same general information for effect size was obtained, though a higher level of agreement (correlation) would be desirable. We did not have the disaggregated NRP effect sizes, that is, Appendix G reports effect sizes aggregated by the outcomes classification. For the most part, it was not possible to compare specific effect sizes directly.

Level of Phonics Instruction

We computed the overall average *d* in a different way than the NRP analysts who first computed an average for each cohort, and then computed a weighted average of these (i.e., the cohort averages) across studies. We obtained averages directly from our database, using "equal study representation" and "compromise" weighting. The analysis of central interest is the difference between systematic and less systematic phonics, since the latter is what many, if not most, students already receive. We used the TP rubric variable (scale 0–2) to describe the level of phonics as a break variable for computing the weighted means given in Table 5. The group labeled "None/not given" in Table 5 (TP = 0) contains treatments that were included as alternatives to systematic phonics, including language-based approaches. These treatments were either not coded by the NRP analysts, or they were used as controls. In the present study, these were coded as treatments if a separate untreated group was available as a control. In other words, our "treatments" consist of both phonics and language-based interventions.

Table 5
Breakdown of Effect Sizes by Type of Phonics Delivered in the Treatment Group*

	Outcome Set

*Note: Both the compromise (*WGT3*) and equal representation (*WGT1*) outcomes sets are given with sums of weights rather than *n*. However, for the *WGT1* set the sums of weights are equivalent to the number of studies. All dependent variable categories are included.

A first approximation of the efficacy of systematic phonics is thus given in Table 5 as the difference between systematic (TP = 2) and less systematic (TP = 1) phonics for which we obtained *d* = .514 - .243 = .27, using *WGT1*. This is about 30% smaller than the magnitude of the effect reported by the NRP. Table 5 also contains results for *WGT2*; however, the results are similar for both sets of outcomes. In the next section, we adjust this effect for other moderators that are correlated with the treatment variable.

Moderator Analysis

As noted above, we created some moderators and borrowed others from the NRP study Appendix G. We examined the 15 moderators below, recognizing that a single outcome could have multiple influences. For this reason, we used weighted multiple regression analysis to sort out the unique contributions of moderators in predicting effects sizes. In this analysis, we examined two sets of moderators:

Set I	**Variable Name**	**Set II**	**Variable Name**
Experimental			
Phonics	TP	Tutoring	Tutor
Language	TL	Duration	Months
Basal	TB	Standardized Test	Standard
Replacement	TR	True Experiment	Random
Control		Grade	Grade
Phonics	CP	Normal v. At Risk/LD	Normal
Language	CL	Expanded v. NRP DV categories	Tag
Basal	CB		
Replacement	CR		

Set I contains the treatment moderators based on the rubric codings of each comparison group. Set II contains other aspects of treatment including tutoring (yes or no); treatment duration; whether the instrument was standardized; whether the experiment was randomized (yes or no); grade; reader ability category; and whether the outcome fell into one of the original 7 NRP categories (yes or no).

We conducted the regression analyses with two orthogonal contrasts for degree of phonics instruction, because effect sizes may not be linear across the categories of TP as suggested in Table 5. The contrasts TP1 and TP2 were coded as:

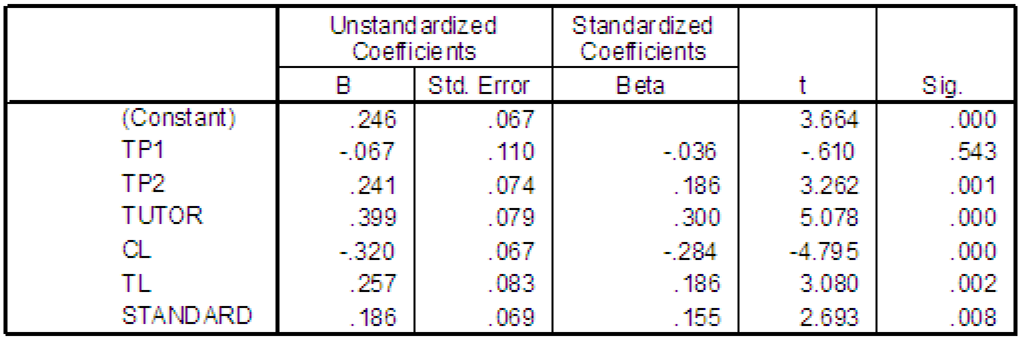
TP1:				
	TP1 = 2/3	if	TP = 0	*No Phonics or Unknown*

	TP1 = -1/3	if	TP = 1, 2	*Some or Systematic Phonics*
TP2:				
	TP2 = 0	if	TP = 0	*No Phonics or Unknown*
	TP2 = -.5	if	TP = 1	*Some Phonics*
	TP2 = .5	if	TP = 2	*Systematic Phonics*

According to this coding, TP1 represents the difference between treatments coded as having no phonics or unknown, on the one hand, and treatments coded as having at least some phonics, on the other. The contrast TP2 represents the specific difference between treatments coded as having some phonics, and treatments having systematic phonics.

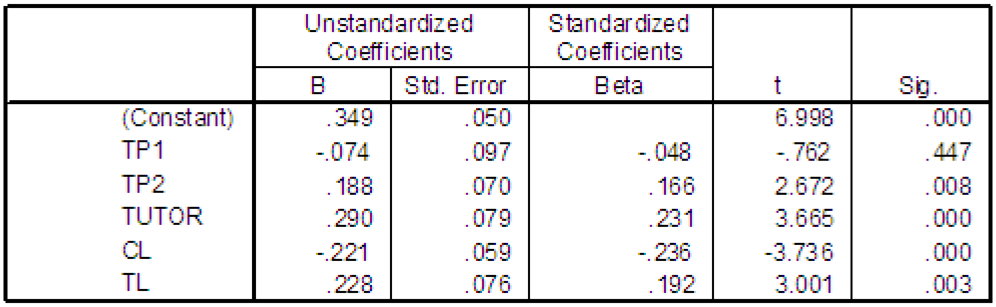
In the regression analyses below, we first entered into the equation the degree of treatment phonics (TP1 and TP2). We then entered the rest of the 14 (7 Set I + 7 Set II) variables into the regression using a forward stepwise procedure. (Note 20) We viewed this as a kind of natural competition of the variables in explaining the results, especially because we took an agnostic stance with respect to reading theory and the previous NRP results. Results for two separate regressions are reported below. In Table 6, regression coefficients are given for the *WGT1* weighting method, and in Table 7 for the *WGT3* weighting method.

Table 6
Regression Coefficients for the Analysis Weighted by *WGT*1, with R^2=.322



	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	.246	.067		3.664	.000
TP1	-.067	.110	-.036	-.610	.543
TP2	.241	.074	.186	3.262	.001
TUTOR	.399	.079	.300	5.078	.000
CL	-.320	.067	-.284	-4.795	.000
TL	.257	.083	.186	3.080	.002
STANDARD	.186	.069	.155	2.693	.008

Table 7
Regression Coefficients for the Analysis Weighted by *WGT*3, with R^2=.199



	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	.349	.050		6.998	.000
TP1	-.074	.097	-.048	-.762	.447
TP2	.188	.070	.166	2.672	.008
TUTOR	.290	.079	.231	3.665	.000
CL	-.221	.059	-.236	-3.736	.000
TL	.228	.076	.192	3.001	.003

For the *WGT1* outcome analysis, the effect of TP1 was $d = -.067$. This means that treatments using no phonics or an unknown degree of phonics had less of an effect than programs that did use a measurable amount of phonics. As shown in Table 6, programs using systematic phonics instruction outperformed programs using less systematic phonics with $d = .241$. The systematic phonics effect, however, is smaller than the effect for individual tutoring ($d = .399$). In addition, standardized tests tended to give larger effects ($d = .186$); studies in which control groups used language approaches had lower effect sizes ($d = ^{-}.320$); and treatments that used language approaches had larger effect sizes ($d = .257$). Results for *WGT3* analysis are given in Table 7. The results are similar to those in Table 6 with the effect for systematic phonics given as $d = .188$; the tutoring effect was moderately smaller ($d = .290$); and the language effects were roughly similar for CL ($d = -.221$) and TL ($d = .228$). (Note 21) Neither analysis provided evidence that randomized experiments give different results than quasi-experimental studies, or that the results differed for the NRP and the expanded set of outcomes categories.

The result for tutoring requires some discussion since it appears inconsistent with the NRP results. The unweighted effect of tutoring $d = 1.09$ is reported in Ehri et al. (2001, Table 2), while the effects for small group and class instruction are given as .44 and .37, respectively. Thus, the unweighted tutoring effect was documented by the NRP. When studies were weighted by size, Ehri et al. (2001, Table 1) the effect sizes for tutoring, small group, and class instruction were .57, .43, and .39, respectively. This change in NRP estimates results from the weighting scheme used, but also from the deletion of the study by Tunmer and Hoover (1993). (Note 22) In the present analysis, the deletion of this study results in a tutoring estimate of $d = .21$ ($p < .007$) while the phonics estimate is virtually unchanged.

The Tunmer and Hoover (1993) study also illustrates an important issue for interpreting the regression results. Recall that there were two treatment groups, and one untreated control group. The first treatment was the Standard Reading Recovery (SRR). It was modified by one and only one change: a systematic phonics component was added. This modified treatment was then given to the second experimental group (MRR). We coded the first group (SRR) as TP = 1 and the second (MRR) as TP = 2, recognizing the difference between the two as the best estimate of the systematic phonics effect. This is what the contrast TP2 represents. The difference between the untreated control group and the phonics groups (SRR and MRR) is the effect estimated by the first contrast TP1.

We examined residuals for the weighted regression analysis and found evidence of one outlier (standardized residual $|z| > 4.5$). This case was removed; however, this decision had very little effect on the model estimates.

Differences Between Outcome Categories

We did not explicitly examine outcomes for dependent variable categories because there were relatively few studies that contributed to any particular category. Our primary goal was to replicate results on the overall efficacy of phonics instruction. However, we did examine residuals from the weighted regression model and test for residual differences between the DV categories given in Table 1. Using an unweighted analysis (to increase *n*) and comparison as the unit of analysis, we found no significant differences, $F(11, 212) = .805$, $p = .635$. This implies that there were no differential effects by DV category. In particular, the average residual for Spelling was virtually zero. We do not think this result implies that phonics instruction is equally effective for all dependent variable categories, but rather that fine-grain discriminations between different types of reading outcomes require more precise data than were obtained from the phonics instruction studies.

Effects by Grade, Unit of Instruction, and Duration

We had a special interest in examining variation in effect size by grade/age. This scatter plot is given in Figure 2, in which it can be seen that in early grades systematic phonics instruction outperforms typical phonics or no/unknown phonics instruction. However, differences among these categories are small shortly after grade 3. A conservative reading of this evidence would indicate that there is no evidence that systematic phonics instruction outperforms alternative treatments after grade 3. However, the phonics indicator is confounded with other treatment variables in the early grades, and the strongest inferences about the efficacy of phonics instruction are obtained from the regression analyses. It should be kept in mind that the trends represent changes in phonics outcomes rather than changes in reading comprehension. The outcomes in Figure 2 appear to have an upward trend beginning just after grade 3. However, the existence of this trend was not verified in the regression analyses using a quadratic term for grade/age. Thus, the information in Figure 2 should be interpreted with some caution.

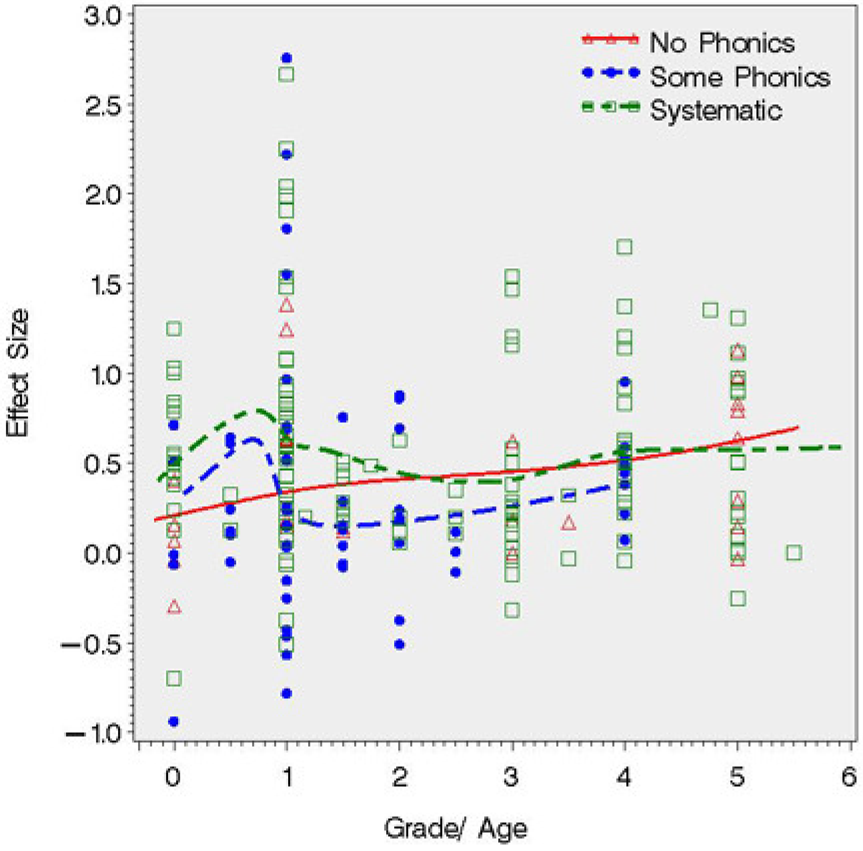


Figure 2. Effect size plotted by grade and degree of phonics instruction.
(On the horizontal axis, the point 0 (zero) represents kindergarten.)

We also plotted tutoring versus other treatment units (i.e., small group and class) in Figure 3. Here it can be seen that tutoring outperforms other instructional unit sizes across the approximate range of kindergarten to fifth grade. Furthermore, there is a suggestion, that tutoring has a greater effect in kindergarten and first grade, but also begins to increase again after third grade.

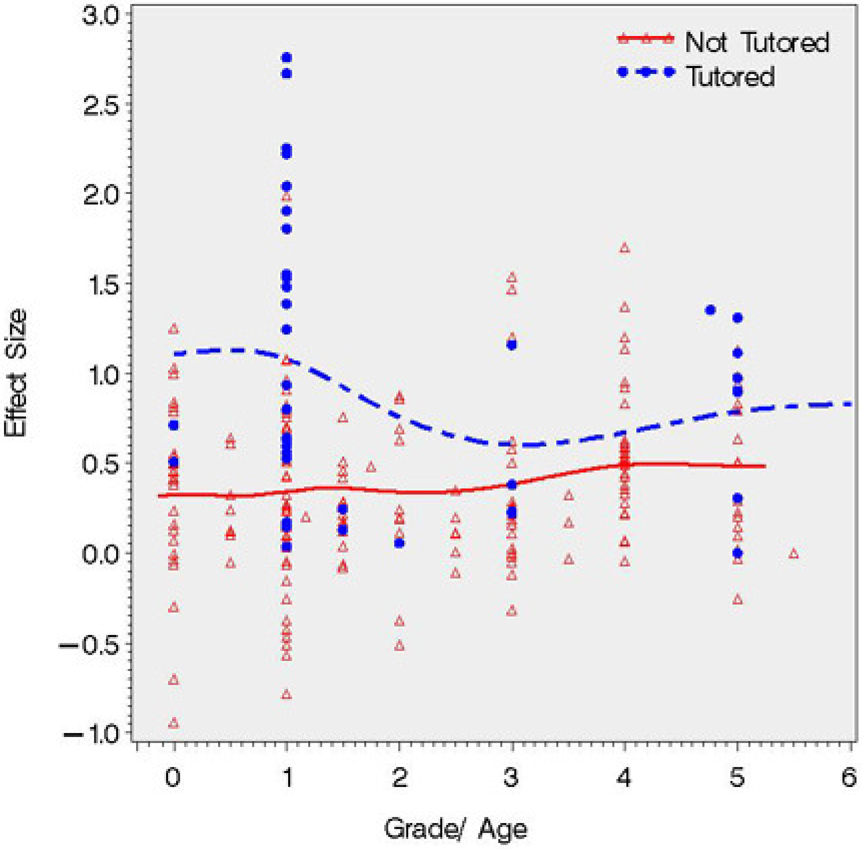


Figure 3. Effect sizes plotted by grade and unit of instruction.
(On the horizontal axis, the point 0 (zero) represents kindergarten.)

In Figure 4 effect size is plotted against the duration of treatment in months. Again the effects of tutoring are superior to those of other units of instruction, but here the effects peak at about 4 months and decline thereafter. We note that duration here denotes the chronological length of treatment and does not indicate intensity (e.g., minutes per day).

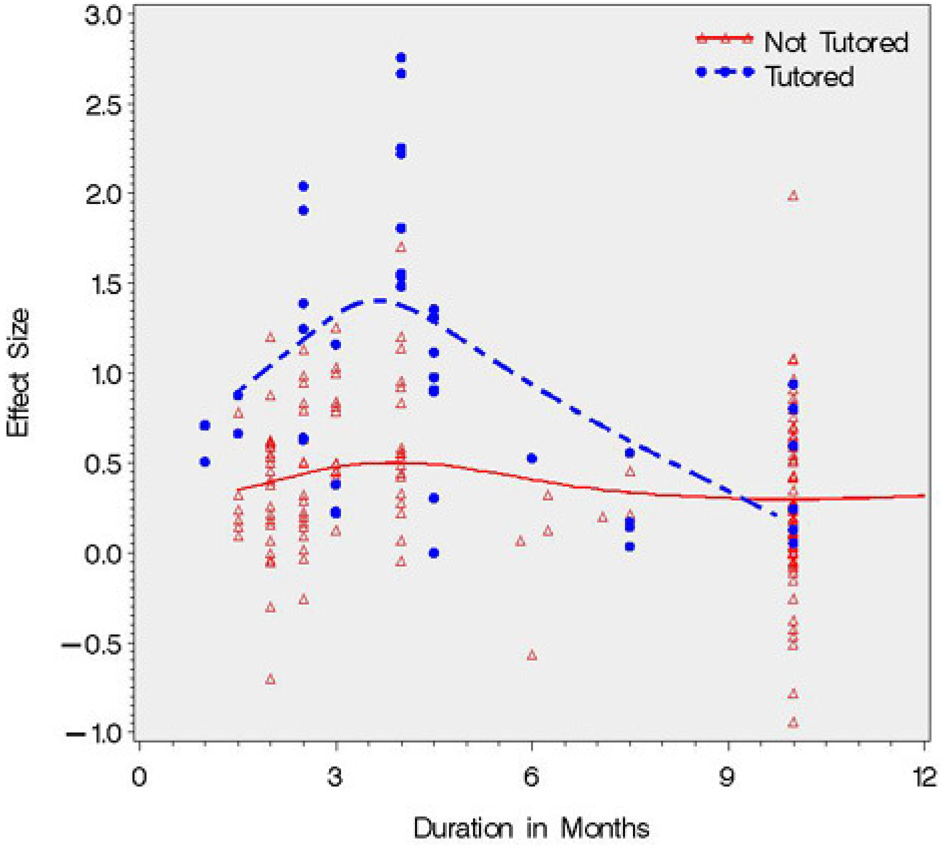


Figure 4. Effect sizes plotted by duration of treatment in months and unit of instruction.

V. Re-analysis: Discussion

Cohen (1988) is commonly cited as suggesting that an effect size of .2 is small, .5 is moderate, and .8 or above is large. However, the primary criterion for judging an effect size in educational research is its potential value for informing or benefiting educational practice. Small effect sizes can be valuable, and likewise large effect sizes can be trivial depending on the treatment and outcome in question. McCartney and Rosenthal (2000) wrote, "There are no easy conventions for determining practical importance. Just as children are best understood in context, so are effect sizes" (p. 175). Average effect sizes only provide information about whether a program works in a general sense. “A more useful question is under what circumstances do programs work best?” (McCartney and Dearing, 2002). To discover these circumstances requires that program characteristics be coded and related to effect sizes. An average effect size can also be evaluated with respect to other kinds of educational treatments. While this information does not provide a definitive rule, it does allow readers to make up their own minds about the practical significance.

Table 8
Average Effect Sizes for Instructional Methods Given by Hattie (1999)

Teaching methods	***d***	**n of *ds***
Direct instruction	.82	253

In the present reanalysis, the estimated effect size for *systematic phonics* was *d* = .241/.188 (for *WGT1* and *WGT3*). This can be compared to effect sizes reported by Hattie (1999, Table 7) for various instructional methods (See Table 8). The overall average is about .4. In

Remediation/feedback	.65	146
Class environment	.56	921
Peer tutoring	.50	125
Mastery learning	.50	104
Homework	.43	110
Teacher Style	.42	*
Questioning	.41	134
Advance organisers	.37	387
Simulation & games	.34	111
Computer-assisted instruction	.31	566
Instructional media	.30	4421
Testing	.30	1817
Programmed instruction	.18	220
Audio-visual aids	.16	6060
Individualisation	.14	630
Behavioural objectives	.12	111
Team teaching	.06	41
* Not given.		

addition, Lipsey and Wilson (1993) examined 302 meta-analyses of a variety of psychological, educational, and behavioral interventions. Interestingly, they also found that the average treatment effect (averaging across meta-analyses) for high quality studies was .4. The largest *d*s in the present study were for tutoring (.399/.290); and use of language activities (about .288/.224). In this context, we would conclude that the advantage of systematic phonics instruction over some phonics instruction is significant, but cannot be clearly prioritized over other influences on reading skills. The regression model suggests, furthermore, that the effects of phonics, tutoring and language activities are *additive*.

It could be argued that the systematic phonics effect is actually larger than the estimate *d* = .241, and so the magnitude of the NRP estimate (about .4) is not an unreasonable expectation. However, the studies examined in this meta-analysis typically did not accurately describe the degree of phonics in the control groups. Thus, while the expectation of *d* = .4 may be plausible, it is not supported by the data. The effect size *d* = -.067 (*p* > .05) for present v. absent/unknown phonics instruction provides a cryptic message regarding alternative approaches to reading instruction. This effect is difficult to interpret because it depends on the "unknown components" of instruction. In the current study, we did not analyze this effect further. However, for teachers who currently teach some phonics, the expected benefit from a shift to systematic phonics is *d* = .241/.188. The present reanalysis suggests that tutoring and language activities are at least as effective in promoting phonics-oriented reading as systematic phonics instruction. (Note 23)

Interpretation of the Evidence on Phonics Instruction

The NRP subgroup on phonics instruction concluded that

> Findings provided solid support for the conclusion that systematic phonics instruction makes a more significant contribution to children's growth in reading than do alternative programs providing unsystematic or no phonics instruction. (NRP, 2000b, p. 2-132)

Based on our reanalysis, the evidence provides ambiguous support for this conclusion. Systematic phonics instruction did outperform treatment conditions in which a more typical or moderate level of phonics instruction was provided. But we identified tutoring and language as critical elements of a reading program in addition to phonics. The data suggest that a reading effect size has the potential to *triple* when these elements are added to systematic phonics instruction. This balance of components is critical in the early grades because the data suggest that after about third grade phonics instruction may be less effective. (Note 24) This is more-or-less consistent with the NRP finding that systematic phonics instruction is most effective in the earlier grades (NRP, 2000b, p. 2-133).

The moderator most strongly related to outcome is the unit of instruction. Tutoring showed a strong effect throughout grades 1-6 (little data are available to extrapolate further). Though shorter phonics programs tended to have larger effects, tutoring was also more effective in this instance. In programs of longer duration, the advantage of tutoring dissipated. Regarding research methodology, we found that standardized instruments (which were published and/or normed) tended to show larger effects, contrary to the expectations of the NRP analysts. This finding, however, was not consistent across the two approaches to weighting (*WGT1* and *WGT3*).

Finally, the regression results we obtained with two different approaches to weighting were roughly similar, but the deletion of one case did make a noticeable impact on the estimated effect for tutoring. This is, unfortunately, the result of a relatively small sample for conducting analyses. In this situation, there is not a single correct model for obtaining estimates, but this is not sufficient reason for ignoring the complexities of the data set. Ultimately, this problem should be resolved by examining larger samples of studies.

VI. Meta-analysis and Public Policy

In the first application of meta-analysis to research on the effectiveness of psychotherapy (Glass et al., 1981), the researchers confronted issues about research integration: how to define the population of studies to be synthesized (only published studies, only studies that met a priori standards of rigor?); how to select and measure the aspects of a study to be related to the outcomes of that study; how to classify studies and calculate their effect sizes when the primary researchers failed to report complete evidence; and how to synthesize outcomes when studies report results for varying sets of outcome measures. (Note 25) The resolutions of such questions and issues worked their way into the development of meta-analysis as a methodology that helps social scientists to distill and validate conclusions from a diverse research literature. This accumulation of research findings is not only helpful for settling disputes among researchers, but has become an important method for designing evidence-based public policies.

Meta-analysis would appear to offer great potential for objectivity and even-handedness in the synthesis of research. Prior to the 1970s, research synthesis had been fraught with bias—the reviewer selected studies that favored one perspective and cast others out, typically for *ad hoc* reasons. Because of its balanced approach, meta-analyses might resolve polarizing conflicts by making the fullest use of the research literature. The recent report from the National Reading Panel was likewise motivated in part by the desire to use the best evidence available to guide instruction in reading. Ironically, this effort has stimulated controversy regarding what constitutes evidence as well as sound research procedures.

Meta-analysis is a kind of quality control mechanism in the process of making sense of numerous individual studies. Yet criteria for the validity of a meta-analysis itself must also be considered. Is there is a general schema for producing meta-analyses that encourages the application of new knowledge? In recent years, it has become evident that a more systematic approach to meta-analysis is required in order for its original ideals to be attained. In the sections below, we explore issues of scientific due process that appear necessary for producing high quality meta-analyses, especially in areas of research laden with diverse philosophies. Included in this discussion are procedural standards, assembly of expert panels, and peer review.

Standards for Meta-Analysis

The NRP was directed to employ "rigorous research methodological standards" in carrying out its charge. However, the NRP report included a total of 7 pages (NRP, 2000a, p. 1-5 to p. 1-11) specifically addressing methodological issues (the seventh page in this section consisted of 2 references). Issues particular to phonics instruction were covered in an additional 5 pages (NRP, 2000b, p. 2-107 to p. 2-111). Altogether, less than one page is devoted to data analysis, and this contains one incorrect formula—a reference to the software used to compute the effects sizes is provided (which presumably used the correct formula). An ensuing report of the results by Ehri, Nunes, Stahl and Willows (2001) devoted just over 1 page to methodological issues beyond study selection. Perhaps this lack of attention to analytic issues was because the NRP interpreted "rigorous standards" to mean "rigorous selection criteria" for including studies, but the results of a meta-analysis depend as much on the rigor of the analytic procedures.

We think it is important for policy-oriented meta-analyses to be designed in advance with clear descriptions of basic analytic strategies. For example, the Campbell Collaborative suggests that researchers provide a rationale for why a particular effect size metric was chosen; under what conditions an effect size will be adjusted for bias; how missing data will be handled; and so forth. The Campbell Collaborative has been working on a broader set of criteria for meta-analysis that will play an increasingly important role in establishing the authoritativeness of a research synthesis. (Note 26)

Constituting Panels and Expert Review

Beyond the Campbell Collaborative principles, there would seem to be an important role of due process in selecting committees to guide meta-analyses, especially for meta-analyses that have great potential for influencing teaching practice. The Congressional bills that directed establishment of the National Reading Panel (SB 939, HR 2192) required that

> The Secretary of Education, or the Secretary's designee, and the Director of the National Institute of Child Health and Human Development, or the Director's designee, jointly shall… establish a National Panel on Early Reading Research and Effective Reading Instruction. (3:13-18)

However, the legislation itself provided only two sentences to guide selection of panel members:

> The panel shall be composed of 15 individuals, who are not officers or

> employees of the Federal Government. The panel shall include leading scientists in reading research, representatives of colleges of education, reading teachers, educational administrators, and parents. (4:4-9)

Contrast this with the selection guidelines of the Institute of Medicine (IOM), which is an institutional constituent of the National Academies of Science:

> Committees are the deliberating and authoring bodies for IOM reports, although strict institutional processes must be followed and the peer review process is independent of the committee. Most committees are consensus committees, meaning the process is designed to reach consensus on the evidence base and its implications. Where the published data are insufficient to support a conclusion, the committee may use its collective knowledge to argue for conclusions. The committee is formed by identifying the expertise and perspectives necessary to address the study topic, soliciting and receiving nominations for candidates from a wide and extensive number of sources, presenting a proposed slate and alternatives to the IOM leadership group, receiving approval from the IOM President, and formally requesting appointment from the NRC chairman. A process of seeking to identify biases and potential conflicts of interest takes place and may disqualify individuals. (Note 27)

The NICHD and Secretary of Education appear to have conducted a selection process consistent with the IOM guidelines in constituting the NRP (Note 28); however, there is no detailed description of the procedure used to choose panelists from about 300 nominees.

Visible selection procedures are important for establishing the perception of balance—that is, a diversity of theoretical and methodological perspectives—as well as actual balance. An appropriate mix of talent may facilitate a knowledge base that furthers dissemination of research findings and improves the design of new research studies. In this regard, the NRP would have benefited by formal inclusion of one or more methodologists. (Note 29) Alternatively, the research would have benefited from an officially appointed group of expert methodologists charged with translating the NRP's oversight into technically rigorous guidelines for design as well as data collection and analysis.

We could not find a description of how independent expert review of the final report was conducted. (Note 30) Moreover, a number of inconsistencies exist between the official Summary (26 pages in length) of the report and the report itself (Shanahan, 2001). If *Teaching Children to Read* had been subjected to a more scrupulous review prior to release, it would have had more potential to command a consensus. We acknowledge the severe time constraints under which the report was produced. However, the role of independent review is to verify and tighten the connections between evidence and summary conclusions. This process is intended to screen out precisely the kinds of inconsistencies and ambiguities that appear in the NRP documents.

VII. Conclusions

The impact of meta-analysis is strongly affected by two design decisions. First, the scientific due process for producing a study is critical to its acceptance. How experts are assembled and provided with resources is as important as their charge. Secondly, the science itself is important. There is no single prescription for producing meta-analyses, even though standards exist for general guidance. In spite of the expertise of research teams, time, and resources available, variability among methodological approaches is probable. Meta-analyses designed to answer controversial questions must anticipate and address this concern. One strategy might be to assemble two different teams of analysts at the onset of a study, each carrying out the five steps of meta-analysis. Another possibility may be to require methods for cross-validation in proposals in response to a formal RFP (request for proposal). Of course, such elaborate procedures are not necessary for all meta-analyses. Rather, they are most relevant to those that affect critical policy decisions, such as the studies conducted by the NRP. In any case, experts (both substantive and methodological) who do not participate in a study should provide peer review. (Note 31)

Meta-analysis is an effective method of "reading" the literature. Yet for many studies in the NRP database on phonics instruction, often little detail was given regarding treatment implementation. The NRP analysts struggled with this issue as evidenced by the number of missing study descriptors in Appendix G. Without careful description of the treatments, their implementation, and the populations of students served, it is doubtful that positive treatment effects can be understood well enough to disseminate to teachers. And without such description, it may be impossible to understand why some treatments do not work as expected. Rigorous qualitative work in reading, which the NRP is currently addressing (Manzo, 2003), has much potential to provide an effective link between theory development, program implementation, and quantitative research findings.

This reanalysis points to a number of moderator variables that may play a prominent role in designing phonics instruction. Obviously, two treatments nominally described as phonics and whole language cannot be directly compared if one uses classroom instruction while the other employs tutoring. We used regression analysis to sort out the effects of moderator variables. This provides an improvement to the one-variable breakdowns used in the NRP report. Based on the regression approach, we found that tutoring and language-based reading activities had effects at least as large as systematic phonics. In addition, the data suggest these effects are additive. These results are starkly different from the quantitative results presented in *Teaching Children to Read*, but interestingly, they are very consistent with two conclusions:

> Programs that focus too much on the teaching of letter-sounds relations and not enough on putting them to use are unlikely to be very effective. In implementing systematic phonics instruction, educators must keep the *end* [original emphasis] in mind and insure that children understand the purpose of learning letter-sounds and are able to apply their skills in their daily reading and writing activities. (NRP, 2000b, p. 2-96).
>
> Finally, it is important to emphasize that systematic phonics instruction should be integrated with other reading instruction to create a balanced reading program. Phonics instruction is never a total reading program. (NRP, 2000b, p. 2-97).

Despite the manifest consistency of these conclusions with the findings of the present report, the ideal role of meta-analysis—to solve controversial issues and thus to improve educational practices—was not directly fulfilled. Two independent teams of researchers arrived at substantially different interpretations of the *same* evidence.

If the NRP results are taken to mean that effective instruction in reading should focus on phonics to the exclusion of other curricular activities, instructional policies are likely to be misdirected. This interpretation of the data results from a design in which simultaneous influences on reading interventions were not adequately coded and analyzed. In particular, early literacy policies are a timely concern, especially as they are interpreted and applied in the federal Early Reading First Program. Program administrators and teachers need to understand that while "scientifically-based reading research" supports the role of phonics instruction, it also supports a strong language approach that provides individualized instruction. As federal policies are formulated around early literacy curricula and instruction, it is important not to over-emphasize one aspect of a complex process.

In our opinion, a sturdier methodology has potential to improve the estimates of the effect size in *all* substantive areas that the NRP examined. Analyses would also benefit from, indeed may require, a substantially larger sample of studies. In this effort, researchers with substantive, methodological, and classroom experience—as well as time and resources—are necessary to find studies, and to propose and test alternative design strategies. While we applaud the NRP for taking the challenging and difficult first steps in summarizing the extant knowledge on reading instruction, it is clear that more work remains to be done.

Acknowledgement

This research was completed with the generous support of The Pew Charitable Trusts. The opinions expressed in this report are those of the authors and do not necessarily reflect the views of The Pew Charitable Trusts. The authors would also like to acknowledge the contributions of Mary Lee Smith and Joanne Yatvin.

Notes

[1] Results of this study were also reported in Ehri, Nunes, Stahl and Willows (2001), and Ehri, Nunes, Willows, Shuster, Yaghoub-Zadeh, and Shanahan (2001).

[2] Details of this selection process are given in Section III.

[3] Meta-analysis can also be performed with studies that that do not examine treatment interventions (e.g., Hunter and Schmidt, 1990). We do not consider other genres of meta-analysis herein.

[4] Meta-analysis is a labor-intensive research activity. It is common to assemble research teams to facilitate the identification and coding of studies within a reasonable amount of time. However, different coders should record the same study information with a limited margin of error.

[5] Readers are referred to Hunt (1997) for an accessible account of the story of meta-analysis.

[6] The first estimate $d = .41$ is for outcomes at the conclusions of programs. The second estimate $d = .44$ is for end of program *or* end of school year, for programs lasting longer (Ehri et al., 2001, p. 414).

[7] Fletcher and Lyon (1998) wrote "In many studies, the research was designed to evaluate the degree of explicitness required to teach word recognition skills. Instruction in word recognition skills, however, occurs along with opportunities for applications to reading and writing, exposure to literature, and other practices believed to facilitate the development of

reading skills in proficient readers. This reflects one of the oldest observations of any form of teaching or training—a targeted skill cannot be learned without opportunities for practice and application." (pp. 59-60).

8. On p. 2-110 the outcome categories are given, but we could find no rationale for this particular classification.

9. Yatvin (2002) reported that "As time wound down, the effects of insufficient time and support were all too apparent. In October 1999, with a January 31 deadline looming, investigations of many of the priority topics identified by the panel a year earlier had not even begun. One of those topics was phonics, clearly the one of most interest to educational decision makers and to the public. Although the panel felt that such a study should be done, the alphabetics subcommittee, which had not quite finished its review of phonemic awareness, could not take it on at this late date. And so, contrary to the guidelines specified by NICHD at the outset, an outside researcher who had not shared in the panel's journey was commissioned to do the review" (p. 368).

10. These did not include follow up comparisons.

11. Garan (2001) shared Yatvin's concern that the NRP did not use a consistent definition of reading. Garan also criticized the NRP meta-analysis for being limited to a small number of studies and for conceptually dissimilar dependent variables. The latter two points, in our view, are problems common to both meta-analysis and narrative review. The degree to which they limit generalizability varies and cannot be determined a priori.

12. A re-examination that began at the problem formulation stage and proceeded to locating relevant studies would provide a more stringent criterion for replicability. It would also be significantly more costly. Though we skipped these two steps, we would agree that problem formulation and data collection significantly shaped the NRP's study.

13. We excluded follow up comparisons, that is, any measurements taken after post-test measurements were excluded from the analyses.

14. While some studies reported age, others reported grade. We converted all results to an approximate grade metric based on the formula grade = age – 5.

15. This formula does not appear in Cooper and Hedges (1994). See Table 16.2 on p. 237.

16. In this simple case, one divides the class-level effect size by √n.

17. We say "distinct" because each cohort involved different groups of students.

18. The data analysis described on p. 1-10 appears to use total *n*s as weights rather than the inverse variance weights described by Hedges and Olkin (1985) on pp. 86 & 110.

19. In the future "pure" statistical weights might be usefully applied when homogenous subsets of effect sizes are identified.

20. We used a highly conservative approach in the forward stepwise selection of independent variables. We required a *p*-value of .01 (PIN) to enter and a *p*-value of .05 (POUT) for removal.

21. Organized language activities were observed in about 30% of both experimental and control comparisons. Note that effective language activities in the experimental group will make the effect size *larger*, while effective language activities in the control group will make the effect size *smaller*. Thus, the two estimates logically have the opposite sign.

22. The NRP deleted one study (Tunmer and Hoover, 1993) with *d* = 3.71 in obtaining the average effect size for tutoring. The value 3.71 arose as the average of 4 effect sizes for WordID (2.94), Spelling (1.63), Nonwords (1.49), and Oral Reading (8.79). It is obvious that the last effect size is an extreme outlier, and the NRP sensibly deleted this in its computations for tutoring. We surmise that this effect size was properly deleted from other computations. We also deleted this effect size (8.71) from our computations, but we included other effect sizes from this study, which ranged from .96 to 3.18.

23. It is interesting that the effect sizes for experimental and control group language instruction are very nearly the same (taking into account reversed signs), which supports the internal design consistency of the treatment codings.

24. The gap is nearly zero at third grade, but widens somewhat at higher grades. Students in later grades do benefit, but are more likely to represent populations of reading disabled students.

25. Material on the origins of meta-analysis was provided by Mary Lee Smith in a personal communication.

26. The Campbell Collaboration is an emerging international effort that “aims to help people make well-informed decisions by preparing, maintaining, and promoting access to systematic reviews of studies on the effects of social and educational policies and practices.” More information is available at http://www.campbellcollaboration.org.

27. This information is available at
http://www.iom.edu/iom/iomhome.nsf/Pages/IOM+FAQs.

28. “Applicants who had taken strong stands supporting or opposing any particular approaches to reading instruction, or with a financial interest in commercial reading materials, were not considered, according to Duane Alexander, the director of the National Institute of Child Health and Human Development, who helped select the panel” (Manzo, 2000). In addition, panelists could not be employees of the Federal government.

29. Two expert consultants in methodology were introduced to the Panel in late January, 1999. It appears that both were made available to NRP members on an as needed basis. This information is available at
www.nationalreadingpanel.org/NRPAbout/Panel_Meetings/01_21_99.htm. Note that the original deadline for the NRP report was January 31, 1999.

30. There appears to be a collection of documents in which the NRP’s interactions are recorded. We do not know if this archive is available for public examination (see Yatvin, 2002).

31. The Campbell group, referenced above, provides design review as a service. It does not appear to review drafts of final reports.

32. The K-3 NFT group size in the Gersten et al. (1988) study is reported as 45. Official documents give *n* = 21.

33. This model was sponsored by the Southwest Educational Development Laboratory. It stressed a developmental approach geared to children whose primary language was not English. In this approach primary language and cultural background are essential to the learning process.

34. The next five effect sizes are from Camilli (1980). They are covariance adjusted based on a modified linear model that includes a linear selection rule.

35. This model was sponsored by the City University of New York. Rather than didactic methods, direct interaction with other children was the primary method of learning. Instructional games developed skills in the areas of language, reading, and arithmetic.

36. This model was sponsored by the University of Florida. The primary emphasis was on motivating parents, and teaching them to set and attain their children's educational goals. Parents spent time as instructional assistants as well as visiting other FT parents.

37. This model was sponsored by Northeastern Illinois University. Entry language and experience of the children are built upon using a method of language elicitation focusing on the use of oral language in all curriculum areas.

38. Dissertation study, see references.

39. For all reading and spelling outcomes, the amount of growth (linear component) in each class was negatively related to initial PPVT-R standard deviations (using class as the unit of analysis).

References

Camilli, G. (1980). A Reanalysis of the Effect of Follow Through on Cognitive and Affective Development (University of Colorado, Boulder). Dissertation Abstracts International, DAI-A 41/04, p. 1366, Oct 1980.

Coles, G. (2003). *Reading the Naked Truth*. Portsmouth, NH: Heinemann.

Ehri, L., Nunes, S., Willows, D., Schuster, B., Yaghoub-Zadeh, Z., and Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly,* 36, 250-287.

Ehri. L., Nunes, S., Stahl, S., and Willows, D. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*(3), 393-447.

Ehri, L. & Stahl, S. (2001). Beyond the Smoke and Mirrors: Putting Out the Fire. *Phi Delta Kappan*, 83(1), 17-20.

Fletcher, J. M., & Lyon, G. R. (1998). Reading: A research-based approach. In W. M. Evers (Ed.), *What's gone wrong in America's classrooms* (pp. 49-90). Stanford, CA: Hoover Institution Press.

Foorman, B. R., Francis, D. J., Fletcher, J.M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, 90, 1-19.

Foorman, B., Francis, D., Novy, D. & Liberman, D. (1991). How letter-sound instruction mediates progress in first-grade reading and spelling. *Journal of Educational Psychology*, 83(4), 456-469.

Garan, E.M. (2001). Beyond the Smoke and Mirrors. *Phi Delta Kappan*, 82(7), 500-506.

Garan, E.M. (2002). *Resisting reading mandates*. Portsmouth, NH: Heinemann.

Glass, G. V, McGaw, B., & Smith. M.L. (1981). *Meta-Analysis in Social Research*. Beverly Hills: SAGE Publications.

House, E., Glass, G., McLean, L., & Walker, D. (1978). No simple answer: Critique of the FT evaluation. *Harvard Educational Review*, 48(2), 128-160.

Hunt, M. M. *How Science Takes Stock: The Story of Meta-Analysis*. (1997). NY: Russell Sage Foundation.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis*. Newbury Park, CA: SAGE Publications.

Krashen, S. (2000, May 20). Reading Report: One Research's 'Errors and Omissions.' *Education Week*, 19(35), 48-50.

Krashen, S. (2001). More smoke and mirrors: A critique of the National Reading Panel (NRP) report on fluency. *Phi Delta Kappan,* 83(2), 118-22.

Layzer, J., & Goodson, B. (2001). *National Evaluation of Family Support Programs*. Cambridge, MA: Abt Associates, Inc.

Lipsey, M.W. & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48(12), 1181-1209.

Lovett, R., Ransby, M., Hardwick, N., Johns, M., & Donaldson, S. (1989). Can dyslexia be treated? Treatment-specific and generalized treatment effects in dyslexic children's response to remediation. *Brain and Language*, 37, 90-121.

Manzo, K.K. (1998, February 18). New National Reading Panel faulted before it's formed. *Education Week*, 27(23), 18.

Manzo, K.K. (2000, April 19). Reading Panel Urges Phonics For All in K-6. *Education Week*, 19(32), 1 & 14.

Manzo, K.K. (2002, January 30). New Panels to Form to Study Reading Research. *Education Week*, 21(20), 5.

Manzo, K.K. & Hoff, D.J. (February 5, 2003). Federal Influence Over Curriculum Exhibits Growth. *Education Week*, 22(21), 1 & 10 & 11.

McCartney, K. & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, 71, 173-180.

McCartney, K. & Dearing, E. (2002). Evaluating Effect Sizes in the Policy Arena. *The Evaluation Exchange Newsletter*, 8(1), 4 & 7.

National Reading Panel. (2000a). *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*. Washington, D.C.: NICHD.

National Reading Panel (2000b). Alphabetics Part II: Phonics Instruction (Chapter 2) in Report of the National Reading Panel: *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction: Reports of the Subgroups*. Rockville, MD: NICHD Clearinghouse.

Orwin, R. G. (1994). Evaluating coding decisions. Pp. 140-162 in H. Cooper & L.V. Hedges (Eds.). *The handbook of research synthesis.* New York, NY: Russel Sage.

Pressley, M. & Allington, R. (1999). Concluding reflections: What should reading research be the research of. *Issues in Education*, 5(1), 165-175.

Shadish, W.R. & C.K. Haddock (1994). Combining estimates of effect size, pp. 261-281. In Cooper, H. & L.V. Hedges (eds.), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation.

Shanahan, T. (2001). Response to Elaine Garan: Teaching Should be Informed by Research, Not Authoritative Opinion. *Language Arts Journal*, 79(1), 71-72.

Tunmer, W. E., & Hoover, W. A. (1993). Phonological recording skill and beginning reading. *Reading and Writing: An Interdisciplinary Journal*, 5. 161-179.

Vickery, K.S., Reynolds, V.A., & Cochran, S.W. (1987). Multisensory teaching approach for reading, spelling, and handwriting, Orton-Gillingham based curriculum, in a public school setting. *Annals of Dyslexia,* 37, 189-200.

Yatvin, J. (2000). Minority View. In *National Research Panel, Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*, pp. 1-6. Washington, D.C.: NICHD.

Yatvin, J. (2002). Babes in the Woods: The Wanderings of the National Reading Panel. *Phi Delta Kappan*, *83*(5), 364-369.

About the Authors

Gregory Camilli
Rutgers University
10 Seminary Place
New Brunswick, NJ 08901

732.932.7496 X8350
camilli@rci.rutgers.edu

Gregory Camilli is Professor in the Rutgers Graduate School of Education. His interests include measurement, program evaluation, and policy issues regarding student assessment. Dr. Camilli teaches courses in statistics and psychometrics, structural equation modeling, and meta-analysis. His current research interests include school factors in mathematics achievement, technical and validity issues in high-stakes assessment, and the use of evidence in determining instructional policies.

Sadako Vargas
As Assistant Professor at Kean University, and Adjunct Professor at Touro College and Seton Hall University, Sadako Vargas has taught in the areas of research methods and occupational therapy. Her interests lie in the use of meta-analysis for investigating intervention effects in the area of rehabilitation specifically related to pediatrics and occupational therapy intervention.

Michele Yurecko
Michele Yurecko is a Ph.D. student in Educational Psychology with a concentration in educational measurement at the Graduate School of Education, Rutgers University. Her academic interests include the study of research methods and design applied to the field of education, and the intersection of educational research, testing and public policy.

Appendix A
Additional Studies

Study ID	
63	Barr, R. (1974). The effect of instruction on pupil reading strategies. *Reading Research Quarterly*, 10, 555-582. This study compared a phonics with a sight word method of instruction. Word learning tasks, word recognition, and comprehension were tested. The process by which subject were assigned to groups was not described, but it was reported that the groups did not differ in age or readiness as measured by the World Learning Tasks. Outcome variables for effect size computation were reported in terms of substitution errors on word reading tasks.
65	Peterson, M.E. & Haines, L.P. (1992). Orthographic analogy training with kindergarten children: Effects on analogy use, phonemic segmentation, and letter-sound knowledge. *Journal of Reading Behavior*, 24, 109-127. This study examined the effect of teaching orthographic analogies based on words that rhyme. Children were tested on segmentation ability, letter-sound knowledge, and reading words by analogy. Subjects were stratified on ability measures, and then assigned by odd and even numbers (sequential ranks) to treatment and control groups.

68	Gillon, G. & Dodd, B. (1997). Enhancing the phonological processing skills of children with specific reading disability. *European Journal of Disorders of Communication*, 32, 67-90. This study compared a 20-hour phonological training program to two groups tested in a previous study published in 1995. We used the original 1995 data in which a group receiving 12-hour phonological training was compared with a group receiving 12-hour semantic syntactic training. Groups were tested with the *Neale Analysis of Reading Ability – Revised*.

Appendix B
Case Studies of Three Selected Studies

Gersten, Darch, & Gleason (1988)

This study used select data from the Follow Through (FT) Planned Variation Experiment, which aimed to increase the achievement and self-concepts of children from economically disadvantaged backgrounds. To give some background, the Follow Through program was intended to pick up where Head Start ended, and maintain presumed academic gains from Kindergarten to third grade. According to White et al. (1973, Volume II)

> [Follow Through] is intended to be a comprehensive project offering educational, medical and dental, nutritional, social, and psychological services to children previously enrolled in Head Start. Follow Through uses a strategy of "planned variation" in approaches to early elementary education, and 20 different models are being implemented in Follow Through sites across the nation. (p. 83).

Fourteen education models (i.e., different treatments) were included in the FT Evaluation (Stebbins et al., 1973), and these varied in the degree of classroom structure, basic skills, and parental involvement. One such model was Direct Instruction (DI), sponsored by the University of Oregon, College of Education. In the DI approach, behavioral methods were used with highly structured teaching materials. Teachers worked with small groups of students, and tests were frequently administered to assess children's progress.

There were two cohorts of students from East Saint Louis, Illinois. Each consisted of a treatment (FT) group receiving DI and Non-Follow Through comparison (NFT) group. One cohort was assessed from grades 1-3 (*n* = 96, 45 for FT, NFT), the other from K-3 (*n* = 56, 21 for FT, NFT). (Note 32) These were the groups providing data for the Gersten et al. (1988) study. Nationally, however, Direct Instruction was implemented at 9 other sites. Outcome measures included the Metropolitan Achievement Test with subtest scores in Word Knowledge, Spelling, Language, and Reading, among others. The NRP analysts choose to compute effect sizes for Reading (*d* = .11, 28) and Spelling (*d* = -.12, .16) for the two cohorts. The Reading effect size was classified as a measure of comprehension. The effect sizes were quite close to calculations from the present study of (.09, 27) for Reading and (-.10, .15) for Spelling. Similar national-level estimates of .14 and .12 for Reading and Spelling (for the K-3 cohort only), respectively, were given by Camilli (1980).

Overall, the results from East Saint Louis are remarkably representative of the national results, but since Direct Instruction was only 1 of 14 other models, we might ask which models showed the largest gains in Reading and Spelling. Camilli (1980) found that two models with the largest Reading effect sizes were Language Development (*d* = .180) (Note 33; Note 34) and Interdependent Learning (*d* = .168) (Note 35). In Spelling, the Parent Education (Note 36) (*d* = .310) and Cultural Linguistic (.341) (Note 37) models had the largest gains. We would add that the Direct Instruction model had the largest gain for MAT Language, Part B (*d* = .327), in which a student was required to recognize asking, telling, and incomplete sentences.

In conclusion, our statistical results are close, in this case, to those of the NRP analysts. Thus, our extended analysis of the Gersten et al. (1988) study can be taken as validation of the consistency of their methodology. However, this case study points to other aspects of the NRP study in terms of its generalizability, or external validity. It is ironic that a single study can strengthen conclusions regarding the value of phonics instruction, and yet the study was originally embedded in a larger study that provided mixed findings with regard to treatment efficacy. Though it is true that the basic skills models (Direct Instruction and Behavior Analysis) had the largest overall gains in the Follow Through experiment, the Direct Instruction model did not outperform other models for Reading or Spelling.

Data from FT models other than Direct Instruction were not included in the phonics instruction meta-analysis for several probable reasons. First, it is doubtful that reports such as those by House et al. (1978) would be identified with the NRP key word searches. It would be virtually impossible in a meta-analysis to anticipate such studies without direct knowledge of their existence. Studies like Camilli (1980) (Note 38) or the FT evaluation reports (e.g., Stebbins et al.., 1977) would not be included because they do not appear in refereed journals. However, even if such studies were located and included, a dilemma would arise because both the NFT and other FT models could serve as controls. Only if enough information were reported for comparing the level of phonics instruction in the alternative treatments could a consistent decision be made. This might be possible even though the data are about 30 years old, but such an in-depth analysis would not be economically feasible.

Tunmer, W. E., & Hoover, W. A. (1993)

This study compared the effects of three different language programs on beginning readers who had been identified as having reading difficulties. Two types of Reading Recovery programs were used for the treatment groups, and the standard intervention program was used for the control.

The first treatment group was the Standard Reading Recovery (SRR) program, which is a remedial reading program developed in New Zealand to “reduce the number of children with reading and writing difficulties.” At risk children were selected and provided with 30-40 minutes per day of individual instruction by a trained teacher for a period of 12-20 weeks. Reading Recovery lessons followed the procedures developed by Clay (1985) and usually included seven activities, one of which was writing a story the child had created. Writing exercises employed phonological awareness training techniques to isolate individual sounds in familiar printed words. Incidental word analysis activities that arose from the children’s responses were available after the children mastered letter identification. This instruction was given in addition to the children’s regular classroom activities.

The second treatment group was the Modified Reading Recovery (MRR) program. It held the parameters of the standard program constant and then added explicit and systematic instruction in phonological recoding skills to the letter identification activities of the standard Reading Recovery program. The control group was the Standard Intervention Group. It received support services that were normally available to at risk readers, mostly funded by the (then) Chapter 1 program. Children were instructed in small groups, and instructional techniques varied greatly and included word analysis activities.

First graders with mean age of 6 years 2 months at the beginning of the school year were drawn from a pool of at risk readers from 30 schools across 13 school districts. The lowest ranked children from each school were given the Diagnostic Survey and Dolch Word Recognition tests. Three matched groups were formed from those who performed at the lowest levels on these tests. The 64 children in the two Reading Recovery treatment groups were drawn from 34 classrooms from 23 schools. The control group of 32 students was drawn from 13 classrooms in 7 schools. Classrooms were "roughly" matched on location, SES and type of classroom reading program. No significant differences were observed between the means of the three comparison groups for age and all pre-treatment measures. The study also reports that two additional control groups of 32 children each were added (p. 170), but there is no further mention of these latter groups.

For this study, the NRP analysts choose two groups, the MRR group and the Standard Intervention group. Effect sizes were then computed for 4 outcome categories: Word ID (*d* = 2.94), Spelling (*d* = 1.63), Nonwords (*d* = 1.49), and Oral Reading (*d* = 8.79). These effect sizes, especially the latter, seem very large, and this could be taken to mean that the effects of systematic phonics instruction were quite impressive. However, it should be noted that systematic phonics instruction was the key element in the MRR group that distinguished it from SRR. By comparing these two groups, we can obtain an estimate of how much improvement resulted from this modification to the standard program. We calculated these effect sizes as Word ID (*d* = -.12), Spelling (*d* = -.25), Nonwords (*d* = -.12), and Oral Reading (*d* = .12). These results indicate that these two groups performed at very similar levels.

The large SRR effect sizes may be due to either the size of the treatment unit or the RR treatment itself, but these two factors are completely confounded in this study. While the children in both Modified and Standard Reading Recovery groups received one-to-one tutoring, the children in the Standard Intervention group received small group treatment. In fact, the authors warned that

> It is important to note, however, that the highly significant results in favor of the two Reading Recovery groups over the standard intervention may not have been due to the Reading Recovery program per se (i.e., the diagnostic procedures, the format of the Reading Recovery lessons, the procedures for discontinuation) but rather to the manner in which the instruction was delivered. Reading Recovery involved one-to-one instruction, whereas the standard intervention involved instruction in small groups. (pp. 172-173)

It is arguable, in fact, that taking the authors' wisdom into account would result in an effect size for Oral Reading of *d* = .12 in contrast the NRP estimate of *d* = 8.79. Once again, we see that there is a significant issue involved in determining the definition of "control group." Whereas the NRP guidelines clearly designate the standard intervention as having the least systematic phonics instruction, it is the comparison of the MRR and SRR groups that is most germane to estimating the systematic phonics effect (in our study represented as the TP2 contrast).

Foorman, B., Francis, D., Novy, D. & Liberman, D. (1991)

This study explored the relationship among phonemic segmentation, word reading and spelling, with the intention of demonstrating the superiority of a more letter-sound (labeled "More-LS") approach of reading instruction. Children receiving less letter-sound instruction (labeled "Less-LS") were not expected to exhibit regularity effects in word reading to the same extent or at the same rate as children receiving More-LS instruction.

Two groups were selected to participate in this study. The Less-LS group was comprised of 40 students enrolled in three first grade classrooms in a Houston, Texas, public school. The More-LS group was comprised of 40 students in three first grade classrooms in two Houston parochial schools. Students in all six classes received one hour of reading instruction daily, and both groups used a basal reading series. Children enrolled in the parochial schools were younger by about 2 months on average ($p < .05$); and they had higher initial reading and PPVT (Peabody Picture Vocabulary) scores, though the latter differences were not significant. Public school classes had, on average, a PPVT standard deviation about 60% larger than that of parochial school classes.

Neither the treatment nor the control regimen was designed or manipulated by the researchers; both reflected the regular teaching habits of the individual classroom teachers. Teachers in the three public school classrooms were described as being committed to "dealing with whole words in meaningful contexts," and described themselves as using a "language experience" strategy to teach reading. The Less-LS teachers used daily story selections from the basal series *Harcourt Brace Jovanovitch Reading* to provide a theme around which instruction was based. Teachers in the three parochial school classrooms were described as being "committed to letter-sound correspondences and having children segment and blend sounds in isolation." (p. 458). Rules for relating letters and sounds, and sequenced spelling patterns were taught using *Scott, Forseman Reading, Phonics Practice Readers, Series B* and *Modern Curriculum Press Phonic Program* (a workbook). Approximately 45 of the 60 minutes devoted to reading instruction were spent on letter-sound activities. A *Scott Forseman* basal reading series was also used.

The study was approximately ten months in duration. Students were administered pre-test measures in October of first grade, with post-test measures administered the following February and May. The following tests were administered: Gates-MacGinitie Reading Test, Basic R, Form 1; Peabody Picture Vocabulary Test – Revised, Form L; a spelling test (researcher-made test consisting of 40 regular and 20 exception words), a word reading test (researcher-made test consisting of 40 regular and 20 exception words); and the 13 item Test of Auditory Analysis Skills, TAAS. There were no significant posttest group differences in TAAS mean scores or trends. There were significant differences in trends of spelling scores (both regular and exception words) and trends of word reading scores (both regular and exception words) favoring the more LS-group. In other words, the more LS-group appeared to improve at a faster rate than the Less-LS group in word reading and spelling.

For the three primary outcome variables (Word Reading, Spelling and TAAS), the researchers did not report standard deviations. In this instance, it appears that the NRP analysts used the simple standard deviation (for the effect size denominator) of class means. According to standard statistical theory, this results in an effect size that is too large by a factor of $\sqrt{n}$, where n is the number of students in the classes. The NRP effect sizes for Word ID ($d = 1.92$), Decoding ($d = 1.67$), and Spelling ($d = 2.21$) are not comparable to those of other studies in which the individual student is the basis for standard deviation calculations. In this case, we converted the effect size to the individual student metric and obtain the following: Word ID (*d* = .48), Decoding (.62), and Spelling (.49). On average, the effect sizes are 3-4 times smaller than those computed by the NRP analysts, which reflect class sizes of about 13 (for participating subjects). Moreover, approximate matching does not completely resolve the issue of what portion of the adjusted *d*s should be attributed to treatment, school type (public versus parochial), and school-by-treatment interaction.

In conclusion, the Foorman study for the most part succeeded at controlling initial differences. However, there is some evidence to suggest that the public school students lagged slightly behind their parochial school counterparts, and that individual differences in ability (PPVT-R) were somewhat larger in the public school classrooms. (Note 39)

We do not know the degree to which this initial difference may have affected posttest differences or rates of growth. However, it is clear that the effect sizes need to be adjusted to the individual student metric.

Appendix C

Description of Vickery, K.S., Reynolds, V.A., & Cochran, S.W. (1987). Multisensory teaching approach for reading, spelling, and handwriting, Orton-Gillingham based curriculum, in a public school setting. *Annals of Dyslexia,* 37, 189-200.

The study reports the results of a four-year study (1978 – 1981) that investigated the effect of the Multisensory Teaching Approach for Reading, Spelling and Handwriting (MTARSH) in both remedial and nonremedial classes in a public school. The study reports the result of California Achievement Test, which were administered annually in April of each year. The MTARSH was developed by adapting the individualized Orton–Gillingham-Stillman method to small homogenous groups of students. The MTARSH employs two basic decoding techniques, synthesizing phonics and memorizing whole words.

The authors report the baseline scores for each grade and the posttest scores of both remedial and nonremedial classes (separately) taken after 1,2, 3 and 4 years MTARSH instruction. The remedial classes were composed of students who qualified for Chapter 1 or special Education/LLD program, at risk of presenting reading difficulties. All other children enrolled in this school were classified as non-remedial. The MTARSH Program was employed for all students, both remedial and non-remedial, in this school (*n* = 426 during the four years covered by this study). The amount of instruction received is equal for both groups- 25-minutes per day for the first graders and 55 minutes of daily instruction for grades 2 through 6. For the remedial classes, MTARSH program was their only instruction in reading, spelling, and cursive writing. The non-remedial classes MTARSH program was taught in lieu of the regular state-adopted spelling and handwriting programs, using the supplemental reading materials and the basal readers. Although detailed instructional method and materials were different in two groups, the MTARSH method used in both classes was treated as comparable in this study.

The baseline score is from the pre-tests administered two years prior to the introduction of the MTARSH program. The intervention effect was measured by the difference between the baseline scores and the posttest scores. The analysis was conducted separately for remedial group and nonremedial groups. The NRP reports eight effect sizes for this study under general reading category. (Alphabetics, Part II. Appendix G. page: 2-174). Effect sizes are reported for 3rd 4th, 5th and 6th grades for both remedial and non-remedial groups, which yielded the 8 effect sizes computed by the NRP team. Through recalculation of the effect sizes using the formula reported (NRP Report, page 1-10) and the sample sizes reported in Appendix G, it was verified that the NRP used baseline averages as the "control group" outcome, and the one-year follow-up test averages as the "experimental" outcome. The effect sizes were reported to represent the magnitude of performance differences between the phonic instruction (Orton–Gillingham method) and regular class instruction that was provided before the MTARSH was instituted. This study examined the effect of *one* instructional method on two different populations; no control group, or other instructional method, was available for comparison. The design is clearly pre-post and does not satisfy a strict interpretation of the quasi-experimental requirement for inclusion (NRP, pp. 1-7 to 1-9).

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Spanish Language Editorial Board

Associate Editor for Spanish Language