

EDUCATION POLICY ANALYSIS ARCHIVES

English Editor: Sherman Dorn
College of Education
University of South Florida

Spanish Editor: Gustavo Fischman
Mary Lou Fulton College of Education
Arizona State University

Volume 16 Number 16

August 30, 2008

ISSN 1068–2341

A Response to Steubing et al., “Effects of Systematic Phonics Instruction are Practically Significant”: The Origin of the National Reading Panel

Gregory Camilli

Rutgers, the State University of New Jersey

Sun Hee Kim

Teachers College, Columbia University

Sadako Vargas

Rutgers, the State University of New Jersey

Citation: Camilli, G., Kim, S. H., & Vargas, S. (2008). A response to Steubing et al., “Effects of systematic phonics instruction are practically significant”: The origins of the National Reading Panel. *Education Policy Analysis Archives*, 16(15). Retrieved [date] from <http://epaa.asu.edu/epaa/v16n15/>.

Abstract

A recent article by Stuebing, Barth, Cirino, Francis and Fletcher critiqued the findings of Camilli, Vargas, and Yurecko (2003) and Camilli, Wolfe, and Smith (2006). With a methodological argument, they attempted to resolve the conflict between these studies and the original report *Teaching Children to Read* (National Reading Panel, 2000). In response, it is argued that three issues must be considered in a fair assessment of the NRP report—program labels or bins, alternative bins, and the role of literacy activities in reading instruction. In this light, three hypotheses ventured by Stuebing et al. are analyzed. It is concluded that the argument by Stuebing et al. does not reveal flaws in the original NRP report by Camilli et al. (2003), though some points of agreement are acknowledged.

Keywords: early reading; phonics instruction; literacy instruction; reading difficulties; National Reading Panel; meta-analysis.



Readers are free to copy, display, and distribute this article, as long as the work is attributed to the author(s) and **Education Policy Analysis Archives (Archivos Analíticos de Políticas Educativas)**, it is distributed for non-commercial purposes only, and no alteration or transformation is made in the work. More details of this Creative Commons license are available at <http://creativecommons.org/licenses/by-nc-nd/2.5/>. All other uses must be approved by the author(s) or **EPAA/AAPE**. **EPAA/AAPE** is published jointly by the Colleges of Education at Arizona State University and the University of South Florida. Articles are indexed by the Directory of Open Access Journals, H.W. Wilson & Co., and SCOPUS.

Una Respuesta a Stuebing et al., "Efectos de Instrucción Fonética Sistemática son Prácticamente Importantes": El Origen del Comité Nacional sobre Lectura

Resumen

Un artículo reciente de Stuebing, Barth, Cirino, Francis y Fletcher critica las conclusiones de Camilli, Vargas, y Yurecko (2003) y Camilli, Wolfe, y Smith (2006). Con un argumento metodológico, estos autores trataron de resolver el conflicto entre estos estudios y el informe original, Enseñar a leer a los niños (Comité Nacional sobre Lectura, 2000). En respuesta, se argumenta que tres cuestiones deben ser considerados en una evaluación justa del CNL: informe-programa de etiquetas o cajas, cajas alternativas, y el papel de la actividades de alfabetización en la instrucción de la lectura. Desde este punto de vista, se analizan tres hipótesis presentadas por Stuebing et al.. Se concluye que el argumento de Stuebing et al. no revela fallas en el informe original CNL de Camilli et al. (2003), aunque se reconocen algunos puntos de acuerdo.

Palabras clave: principios de lectura, instrucción fonética; alfabetización; dificultades en la lectura; Comité Nacional sobre Lectura; meta-análisis.

A recent article by Stuebing, Barth, Cirino, Francis and Fletcher (2008) reviewed the findings of a pair of papers by Camilli and colleagues (Camilli, Vargas, & Yurecko, 2003; Camilli, Wolfe & Smith, 2006). From a methodological perspective, they attempted to resolve the disparity in results of these studies with those of the original report *Teaching Children to Read* (or *TCR*; National Reading Panel, 2000). Stuebing et al. argued that similar estimates of the efficacy of phonics instruction can be obtained from National Reading Panel (NRP) and the Camilli et al. (2003) data sets, provided that same parameter is estimated. As shown below, this argument is artificial and evades key policy questions regarding reading instruction. In addressing these weaknesses, one objective of the current article is to identify inadequacies in the extant research literature that impede effective instructional applications of reading research. These inadequacies also limit the extent to which meta-analytic studies can improve the delivery of reading interventions. A second objective is to identify a point of agreement with Stuebing et al. that may eventually lead bring reading researchers nearer to consensus.

It is helpful, first of all, to understand the original purpose of the NRP meta-analysis. The authors of the NRP report wrote "Does systematic phonics instruction help children learn to read more effectively than nonsystematic phonics instruction or instruction teaching no phonics?" (p. 2-92). The astute reader will observe that two different questions are conflated in this quote: First, does systematic phonics instruction help children learn to read more effectively than *nonsystematic phonics instruction*? Second, does systematic phonics instruction help children learn to read more effectively than *instruction teaching no phonics*? One important difference between these two questions relates to standards for comparison, that is, the baseline against which efficacy is measured. For the first question, it is some degree of phonics instruction, and for the second question, it is *instruction teaching no phonics*. As argued below, the NRP's compound question is

muddled in terms of the report's implications for practice. While the NRP found a positive effect using the comparison *instruction teaching no phonics*, Camilli, Wolfe, and Smith (2006) showed that the effect of systematic phonics instruction is positive but much smaller than *nonsystematic phonics instruction*. Perhaps this is why neither the NRP nor Stuebing et al. (2008) appear to have given much thought to how the effect of systematic phonics instruction generalizes to instructional practice.

While the basis for comparison in the compound question above can be operationalized in an experiment as the presence or absence of phonics instruction in a comparison or control group, an additional claim for translating research results into practice was explicit in NRP's summary of results:

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; emphasis added)

It is important to recognize that shifting the comparative benchmark from no phonics to less phonics instruction to "alternative programs" also has important consequences, which an analysis focused strictly on parameters trivializes. Without addressing how reading research generalizes to the highly diverse contexts in which reading instruction occurs, teachers and instructional leaders are provided insufficient guidance, if not incorrect inferences, for the purpose of implementing educational improvements.

In addition to identifying the correct comparison for meta-analysis, one must consider and recognize the joint influences of different activities on reading outcomes of young children. There may be many such influences or moderators, and Camilli et al. (2003) uncovered important effects for both language-rich activities and tutoring. In their paper, Stuebing et al. (2008) gave some attention to the benefits of such moderators, but they chose not to examine these influences closely. While they acknowledged the possibility that such instructional activities may enhance the effects of phonics instruction, such activities were essentially treated as indirect effects of phonics instruction. Two questions are relevant with respect to this tacit assumption. The first is whether there is sufficient support for the hypothesis that the benefits of language instruction are indirect effects. The second question is whether there is a clear distinction between literacy- and phonics-based reading instruction.

What is the Critical Competitor? Consequences for Meta-analysis

Three methodological concerns essential to evaluating reading programs-instructional benchmarks, alternative program benchmarks, and the role of literacy activities in reading instruction-are discussed in this section. In following sections, we assess the hypotheses of Stuebing et al., which they proposed in their examination of the studies by Camilli et al. (2003, 2006) and Hammill and Swanson (2006). After new analyses of the arguments of Stuebing et al. are provided, the conclusions of Stuebing et al. are discussed and a central point of agreement is acknowledged.

Instructional Benchmarks

An experimental question in educational research typically involves the differential performance of a treatment and a comparison group. In *TCR*, the NRP selected the comparison

group as the one receiving the least phonics instruction (NRP, 2000, p. 2-103). However, the comparisons were more difficult than this simple device might suggest:

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. (p. 2-103)

To address this question, Camilli et al. (2003) coded both treatment *and* control groups by the degree of phonics instruction. It is not clear that Stuebing et al. (2008) fully grasped this distinction in claiming that Camilli et al. (2003) “made all comparisons against a no phonics control” (p. 127). In fact, this type of comparison exists in very few of the studies identified by the NRP: the presence of some phonics instruction in control groups occurred frequently. Consequently, it is less important that a certain group is used as a comparison than that the choice of this group is consistent as possible and coded accurately in terms of its instructional characteristics. Because Stuebing et al. incorrectly assumed that Camilli et al. (2003) chose a no-phonics comparison group for computing effect sizes, the analytic model based on Stuebing et al.’s Figure 1 is incorrect. The effect of this misreading is examined below.

Alternative Program Benchmarks

A more important issue in evaluating the practical significance of an educational intervention is how to choose useful benchmarks for evaluating instruction. The NRP used three benchmarks: “no phonics instruction,” “forms of instruction lacking emphasis on phonics instruction,” and “alternative programs.” But it appears that no distinction was drawn between the three in assessing the benefit of reading outcomes. It would not be surprising to find that the benefit of a phonics program appears larger when compared to standard instruction than when compared to an alternative reading program, because a number of factors (e.g., professional development, improved instructional materials) are associated with a program intervention that are not associated with standard or pre-existing instruction. The NRP analysts and Stuebing et al. embedded these decisions tacitly in their analyses. Yet the three types of comparative information are relevant to instructional choices, and it would seem obvious that analysts should take this factor into account when informing practitioners.

Practitioners need to know the comparative performance of their real-world choices. The idea that informative value is determined comparatively is a well understood principle, especially in the field of program evaluation. Scriven (1981) wrote “It is a principal maxim of product evaluation as expounded here that [evaluation] is very rarely useful unless comparative” (p. 136). He noted that a car’s repair record can be evaluated against a minimally accepted standard, but comparison to the repair records of another car may provide more practical information. Scriven argued that consumers need to know about “critical competitors,” and thus it would be much less informative to compare an Accord with a Lincoln or Jeep than with a Camry. If a critical competitor is not readily available in a particular situation, a useful evaluation would then require a creative alternative. Scriven (1981) added that “evaluation is usually supposed to serve decision making, and decision making is choosing between alternatives, and if evaluation does not look at the comparative merits of the alternatives, it is not serving decision making” (p. 137). As argued by Cook (1997), research literature synthesis shares important goals with a program evaluation because it is often intended to communicate information that impacts decision making.

As in program evaluation, the phonics meta-analysis reported in *Teaching Children to Read* was intended to lead to the intelligent choice of reading programs. Yet given the report's fuzziness in establishing comparative value for phonics instruction, it falls short as a resource for decision makers. As observed by Dawson and Tilley (1997) "Programs are always introduced *into* pre-existing social contexts, and these prevailing social conditions are of crucial importance when it comes to explaining the successes and failures of social programs" (p. 411). Whether a standard for comparison is an organized and intentional activity or exists by *fiat* is a crucial concern to any evaluator. Moreover, given nearly equivalent outcomes, the shrewd evaluator is always on the lookout for a program that is more cost effective than an alternative program.

In providing a more concrete notion of the idea of "critical competitor," Camilli et al. (2003) examined one study in the NRP database by Tunmer and Hoover (1993), which compared the effects of three different reading programs on beginning readers identified as having reading difficulties. Two types of Reading Recovery program 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. The second treatment group was the Modified Reading Recovery (MRR) program, which held the main ingredients of the standard program constant and added explicit and systematic instruction in phonological recoding skills to the letter identification activities of the standard Reading Recovery program. The control group, the Standard Intervention Group, received diverse support services that were normally available to at risk readers.

For this study, the NRP analysts chose the Modified Reading Recovery (MRR) group as the treatment group and Standard Intervention as the comparison 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 ($d = 8.79$). Yet systematic phonics instruction was the key element in the MRR group that distinguished it from Standard Reading Recovery (SRR), and if SRR was taken as the critical competitor, then the effect sizes would be recomputed as Word ID ($d = -0.12$), Spelling ($d = -0.25$), Nonwords ($d = -0.12$), and Oral Reading ($d = 0.12$). The take-home message is that the MRR group performed at a similar level to its critical competitor SRR. Though this is only one example of how choosing a label or bin for comparison can affect computation of effect sizes, it illustrates why Camilli et al. (2003) coded both treatment and comparison groups for instructional activities. If, on the other hand, the critical competitor to the MRR group is taken to be the group with the least amount of intervention, this decision creates a different foundation for generalizing the research findings. Briggs (2007, p. 18) noted that "This is critically important because it is the choice of control group that makes an estimated causal effect interpretable. The control group provides a frame of reference of an effect." This choice should play an explicit role in discussion of how the results of a meta-analysis may be applied in practice. But it is important to recognize that far less description is typically given to comparison group instruction in the research literature, and this omission results in a serious impediment to comparing the value of different approaches to reading instruction.

Estimating the NRP Effect Sizes

Stuebing et al. contended that the NRP report and Camilli et al. (2003) asked different questions of the data and consequently estimated different parameters. They purport to show that if the same parameter were estimated, the results of the two studies would be similar. To answer this question, an argument (to which we refer as the *a-b-c* argument) was given in which three effects were defined:

a = systematic phonics versus "no phonics" control,

b = some phonics versus “no phonics” control,

c = systematic phonics versus some phonics,

and these effects were asserted to follow the relationship $a = b + c$. (Stuebing et al. illustrate this argument in their Figure 1.) From Table 5 of Camilli et al. (2003), they obtained estimates as

$$a = 0.514$$

$$b = 0.243$$

$$c = a - b = 0.271$$

Thus, the average effect size of the systematic phonics is estimated as

$$(a + c)/2 = (0.514 + 0.271)/2 = 0.393.$$

They claim this result is essentially the NRP result (0.39). Accordingly, Stuebing et al. then argued that Camilli et al.’s (2003) results confirm those of the NRP study. As noted above, this interpretation of the data is not consistent with Camilli et al.’s (2003) coding of the data, because effect sizes with which a and b are computed were not necessarily based on a “no phonics” control. Moreover, the average effect sizes given in Camilli et al.’s Table 5 are not pure measures of a and b . For example, effects sizes for treatments coded as *systematic interventions* were sometimes obtained with control groups receiving *some phonics* or *none/not given*. The same is true for effects sizes for treatments coded as *some phonics*. Though $d = 0.39$ is still approximately correct when no-treatment controls are isolated, the more palpable flaw is that these simple differences do not control for literacy activities and tutoring—which have effects at least as large as systematic phonics. Important for practitioners is the fact that tutoring can occur with either phonics or literacy instruction; by itself, it is not a type of reading instruction.

Taking the Stuebing et al. argument at face value for classrooms in which some phonics instruction is already provided, the gain from switching to an instructional approach incorporating systematic phonics would be $c = 0.271$. While the projected gain would be much larger for students switching from no phonics instruction to systematic phonics instruction ($a = 0.514$), it would not serve instructional leaders well to suggest the latter effect would be realized *independently of the pre-existing instruction context*. Though it is doubtlessly true that some teachers use virtually no phonics instruction, it is false to assume that this behavior is widely generalizable. A number of researchers have reported that most teachers think decoding skills are important and include decoding instruction daily in their classrooms (Baumann, Hoffman, Moon, & Duffy-Hester, 1998; Morrow & Tracey, 1997; Pressley, Rankin, & Yokoi, 1996). The NRP (2000) also cited a study of Fisher, Lapp, and Flood (1999) in a survey of 118 California teachers found that 64% of the K through 2 teachers integrated phonics instruction into their lessons (with some extra isolated phonics), and the remainder taught phonics as a separate part of word study. Rather than assuming what is true in a particular context, we think it is important first to delimit the contextual information within studies of the NRP database, and then to propose and defend alternative means of generalizing.

The X-Y-Z Argument

A second issue regarding interpretation of the NRP effect is also problematic. As mentioned above, one could interpret the effect size as the likely benefit of a programmatic intervention compared to standard instruction, or as the relative benefit of one intervention compared to another intervention. In Camilli et al. (2003), an average effect size was provided for a set of organized (or programmatic) treatments in which no specific phonics intervention was given. For this latter set of effect sizes, interventions were often not described well enough to determine how much phonics instruction occurred. Nonetheless, it is clear that these alternative programs in the comparison group emphasized alternatives to phonics instruction, though it is probable that some phonics instruction

occurred (see coding definitions in Table 2 of Camilli et al., 2003). Following this schema rather than the *a-b-c* argument, the effect estimates given by level of instruction in Camilli et al.'s Table 5 are:

$X = 0.514$ = average effect of systematic phonics treatments

$Y = 0.243$ = average effect of some phonics treatments

$Z = 0.356$ = average effect of no/unknown specific phonics treatments

As shaped by the guiding question proposed above of whether phonics interventions fare better than other systematic interventions not emphasizing phonics, the effect would be estimated as $X - Z = 0.158$. Note that the calculations of Stuebing et al. employ *a* rather than the programmatic comparison $X - Z$, and *b* instead of $Y - Z$. If the calculations are revised so that systematic interventions are compared with other forms of intervention, then

$$\frac{(a+c)}{2} \rightarrow \frac{(X-Z)+(X-Y)}{2} = 0.215,$$

which is approximately half the size of the reported NRP effect. Note that with this approach the effects of *some phonics* instruction compared to *no/unknown phonics* instruction becomes negative ($0.243 - 0.356 = -0.113$). This is an indication that treatments incorporating some phonics on average were slightly outperformed by treatments in which phonics was not a major focus.

Conversely, we can also determine from the equation in the above paragraph that the Stuebing et al. estimate of the systematic phonics effect can be expressed as

$$a = 0.514 = 0.215 + \frac{Y+Z}{2}.$$

The last term is the average effect of treatments that are identified as either *some phonics* or *no/unknown phonics*—in other words, Stuebing et al.'s *a* estimate includes the effects of treatments that are outside what they claim is the preferred treatment. Claiming that the effect of systematic phonics is $d = 0.514$ is akin to claiming that the effect $(Y + Z)/2$ was somehow generated by or subordinate to systematic phonics instruction. (Economists would call this an *exogenous* effect unrelated to *systematic phonics instruction*.) This is a strong assumption for which no warrant was provided; indeed, no awareness of this issue can be detected in either the original NRP (2000) report or in the new analysis by Stuebing et al.

Given the above analyses, we conclude that the ambiguities present in NRP statements of purpose or mission (rather than the data analysis *per se*) led to implicit choices in defining program benchmarks for estimating and interpreting effect sizes. If control groups are chosen with untreated children (e.g., children sitting in a room doing silent exercises), then the effect sizes cannot be interpreted consistently with the NRP benchmark “forms of instruction.” A stronger case can be made for interpreting the effect size as a value-added effect representing the difference between a systematic intervention and standard instruction, but what constitutes standard instruction varies from school to school. The basis for the generalizability of the statement $d = 0.514$ has important limitations, and this is what led Camilli et al. (2003) to perform a multiple regression analysis to assess the relative contribution of different factors to reading outcomes.

Finally, if the key comparison is to be made between systematic interventions, then the basis for generalization may be broader, but two problems remain: first, the advantage of systematic phonics to “some phonics” intervention is modest ($d = 0.271$), and second, the advantage of systematic phonics to interventions not emphasizing phonics is quite small ($d = 0.158$). *Value-added* effects should be distinguished from *comparative* effects. While the former is defined as the benefit of an intervention above and beyond that of pre-existing standard instruction within a school, the latter is defined the relative benefit of one organized reading intervention compared to another. Both kinds of information are useful.

Regression Models

Stuebing et al. (2008) also attempted to use the *a-b-c* argument with regression results presented in Table 7 of Camilli et al. (2003). In this analysis, they make the following equivalence:

$$a \equiv \text{intercept} + b_{\text{TP2}} / 2 = 0.349 + 0.188/2 = 0.443,$$

where b_{TP2} is the coefficient tied to the contrasts among *no/unknown phonics*, *some phonics*, and *systematic phonics* (Camilli et al., 2003). Correcting a minor error¹ results in

$$a \equiv \text{intercept} + b_{\text{TP2}} = 0.349 + 0.188 = 0.537.$$

This estimate is very close to $X = 0.514$, which is to say that what Stuebing et al. did in effect was to add an exogenous model intercept to the systematic phonics effect (as the NRP did implicitly before Stuebing et al.). In the language of statistical analysis, the intercept is also known as the model origin, and this is the explanation to the subtitle of the current article. It is more than a play on words, because without the contribution of the intercept, the evidence presented in *Teaching Children to Read* does not provide “solid” support for the claim that phonics systematic instruction has a significant practical effect relative to other programmatic interventions. The NRP combined an estimate arising from unknown factors (an exogenous effect) with an estimate having stronger causal justification, and then claimed this hybrid benefit had policy implications, if not a causal interpretation.

Effect Sizes in Context

Because researchers exert little if any control over exogenous factors, we have argued that one important component of determining instructional benefits should be conceptualized in terms of comparative value. For example, an effect size estimate in this case corresponds to the question “What benefit does a systematic phonics intervention have compared to standard instruction or an alternative reading program?” For an instructional context in which reading instruction is dysfunctional, a different question might be useful for guiding an evaluation. Yet it is likely that many instructional enterprises are currently providing adequate reading instruction and are interested in improvement that might result from adopting a new reading curriculum. In this situation, the switch to an intervention emphasizing systematic phonics might prove to be a disappointment.

Stuebing et al. (p. 125) claimed that “even the critics of the NRP report concede that different studies show small effects significantly favoring systematic phonics” requires a similar qualification. In fact, Camilli et al. (2006) found a statistically nonsignificant effect, and in terms of *practical significance* they found that the benefit of systematic phonics was small. In arriving at this conclusion, Camilli et al. relied on the study by Hattie (1999) who presented average effect sizes for 18 types of teaching methods. Of this set of studies, the effect size $d = 0.123$ from Camilli et al. (2006) for systematic phonics instruction falls at the 11th percentile, and we concluded that the moderator variables of language and tutoring had larger effects. The effect size $d = 0.188$ also falls

¹ Possibly to purify estimates, they deleted 25 contrasts from the database for which treatments did not include systematic or some phonics and then reran the multilevel regression (their Table 1). However, if we assume that other moderators such as language or tutor components also contribute to the reading program effects, deletion of 25 cases solely based on the phonics indicator is as likely to introduce a model specification error as it is to purify estimates. For this reason, we employed the estimates in Table 1 of Camilli et al. (2006) using the full sample for the further analysis. Also, they did not use the orthogonal contrast coding approach employed by Camilli et al. (2006). If orthogonal coding (-0.5, 0.5) is used, and the systematic phonics effect when compared to some phonics should not be divided by a factor of two.

well below the 50th percentile. Thus, systematic phonics instruction has a potentially small return in some situations, and greater value might result from improving other components of reading instruction. Yet even this statement was further elaborated by Camilli et al. (2006), who wrote that systematic phonics instruction given every day for 10–15 minutes may have a benefit that exceeds its cost as an instructional tool for many populations of early readers.

What Is the Intervention? Literacy Instruction as a Mediator

Camilli et al. (2003, 2004, 2006) estimated the effects of literacy instruction including language-rich activities that occurred in some treatment and some comparison groups. Because a number of instructional components were typically combined, estimating the effects of teaching phonics required judgments about what is and what is not systematic phonics instruction. For example, phonics instruction might have been taught as a separate instructional module, or it might have been embedded in a literacy curriculum. In the embedded approach, grapheme-phoneme relations are taught in the context of words and text. The NRP described this type of instruction as having focus on larger subunits of words.

The difficulty of coding treatment emphases within the NRP meta-analysis database is illustrated by two studies. Torgesen, Wagner, Rashotte, Lindamood, Conway and Garvan (1999) evaluated two different types of phonics instruction. As described in the NRP report, one approach to instruction employed in this study was providing “very explicit and intensive instruction,” while another was described as providing “systematic but less explicit instruction.” The latter was characterized as embedded phonics. In the second study, Foorman, Francis, Fletcher, Schatschneider, and Mehta (1998) described an embedded phonics treatment which included “whole-class activities such as shared writing, shared reading, choral or echo reading, and guided reading” (p. 40), and teachers would “frame a word containing the target spelling pattern during a literacy activity” (p. 40). From these studies, we can conclude that the terms *explicit* and *systematic* may not be equivalent; reading instruction may often blur the line between phonics and literacy activities.

The NRP coded phonic instruction by type, as synthetic, large subunit, mixed, and miscellaneous. In contrast to the NRP methods, the reading experts in Camilli et al. (2006) coded studies according to the intensity of language activities and phonics instruction. We recognize that the reading research literature is not well described for the purpose of identifying the unique contributions of different types of instruction, but this is hardly a rationale for ignoring these complexities. If literacy-based activities constitute an important component of reading instruction, then it is extremely important to understand how this component can be combined with others for optimal learning. For this purpose, it is unwarranted to presume that effective literacy activities can be designed as contingencies relative to phonics instruction or that such benefits will accrue from unorganized literacy activities.

Integrative Approaches to Reading Instruction

One purpose of meta-analysis is to discover what components of instruction contribute to positive student outcomes. It is the goal, based on this empirical evidence, to enable practitioners to provide more effective reading instruction by designing an optimal arrangement of instructional

components. An exclusive focus on systematic phonics is insufficient for this purpose.² Using estimates from Table 1 of Camilli et al. (2006), the regression equation for obtaining a projected treatment benefit for a treatment can be written as follows:

$$\hat{d}_T = 0.348 + 0.489\text{Tutor} + 0.123\text{TP2} + 0.210\text{TL2},$$

where the Tutor variable contrasts tutoring with small-group or whole-class instruction, TP2 contrasts systematic versus “some” phonics in the treatment group, while TL2 contrasts systematic versus “some” literacy activities in the treatment group. The corresponding equation for an alternative treatment is

$$\hat{d}_A = 0.348 + 0.489\text{Tutor} + 0.403\text{TL2}.$$

Here, systematic phonics occurs only the treatment groups, but note that tutoring and systematic literacy³ can occur in either group. This alternative approach for projecting treatment and alternative effects is useful because in the NRP database, alternative treatments tended to include the least phonics instruction.⁴ In parallel, the two equations above allow one to examine the benefit of systematic phonics against a critical competitor. Note that the comparison implicitly removes the intercept, which in the NRP database resulted from largely unknown or exogenous factors.

To illustrate this approach, a set of predicted values is given in Table 1 that facilitates a fair comparison among alternative programs for reading instruction. It can be seen by inspecting columns 1 (outcome in the treatment group) and 2 (outcome in the alternative treatment group) that both columns show a benefit; however, the comparative benefit changes sign depending on the particular combination of instructional components. For example, in the third data row of Table 1, the treatment group outperforms the control group (.56 - .35 = 0.21); but in the fourth row, the alternative treatment group outperforms the treatment group (.56 - .75 = -0.19). The disparity can be traced to the presence or absence of systematic literacy in the alternative treatment group. There are two important messages here. First, decisions about reading instruction should not be driven by a single decision regarding the inclusion of systematic phonics. Instruction has to make sense as a package of components. Second, the results of reading research may appear inconsistent if researchers do not describe what is happening instructionally for all groups in a comparative study.

² Stuebing et al. explored various combinations of treatment moderator to search for optimal effects. Though they did not use same procedures employed by Camilli et al. (2006), the estimate of *systematic phonics* (labeled TP2) was obtained $d = 0.183$ ($p = .042$).² Camilli et al. (2003, 2006) reported effects of similar magnitude at $d = 0.188$ ($p > .05$) and $d = 0.123$ ($p > .05$). The Stuebing et al. estimate, whereas the same estimate obtained by Camilli et al. (2006) was nonsignificant at $\alpha = .10$ ($p = .183$).

³ The procedure for coding systematic literacy is detailed in Camilli, Wolfe, and Smith (2006).

⁴ Using the regression estimates from their Table 1, Stuebing et al. calculated the projected effect sizes in their Table 2, but used a negative effect for control group’s systematic literacy component (CL2) in the fourth column. This is a faulty application of the idea that we have referred to elsewhere as “design consistency.” If literacy-oriented instruction is effective, it produces a positive regression estimate in the treatment group, but a negative estimate in the control group. It is important to recognize that the negative estimate in the control group still indicates a beneficial effect because the numerator of an effect size is calculated as “treatment (say A) mean minus control mean (say B).” As moderator variable such as literacy instruction increases performance in the control group, B gets larger and A-B gets smaller. Thus, to predict the effect of literacy-oriented instruction in the control group, we would propose to use the absolute value of the regression coefficient, or $d = 0.403$.

Table 1
Projected effect sizes for treatment groups

| Effect size | | Program components | | | |
|-------------|-------------|--------------------|-------------|-----|----------|
| Treatment | Alternative | Treatment | Alternative | | |
| d_T | d_A | TP2 | TL2 | TL2 | Tutoring |
| 0.35 | 0.35 | no | no | no | no |
| 0.35 | 0.75 | no | no | yes | no |
| 0.56 | 0.35 | no | yes | no | no |
| 0.56 | 0.75 | no | yes | yes | no |
| 0.47 | 0.35 | yes | no | no | no |
| 0.47 | 0.75 | yes | no | yes | no |
| 0.68 | 0.35 | yes | yes | no | no |
| 0.68 | 0.75 | yes | yes | yes | no |
| 0.84 | 0.84 | no | no | no | yes |
| 0.84 | 1.24 | no | no | yes | yes |
| 1.05 | 0.84 | no | yes | no | yes |
| 1.05 | 1.24 | no | yes | yes | yes |
| 0.96 | 0.84 | yes | no | no | yes |
| 0.96 | 1.24 | yes | no | yes | yes |
| 1.17 | 0.84 | yes | yes | no | yes |
| 1.17 | 1.24 | yes | yes | yes | yes |

While Stuebing et al. may not have intended to predict the cumulative effect of program components, this task seems more likely to provide insights for instructional design. If the effects estimated by Camilli et al. (2006) are additive, then the predictions in Table 1 might serve as useful guidelines to policy makers about what advantages might result from combinations of instructional components in newly designed interventions (provided these intervention deliver the same intensity of treatment as in the research studies). The key to using information like that provided in Table 1 is that decision makers must consider carefully the pre-existing instructional ecology in which an intervention will be implemented. That is, only a knowledgeable person “on the ground” can know which row of Table 1 is most relevant for implementing change.

Is Phonics Instruction more Fundamental than Literacy Instruction?

The format of Table 1 assumes that moderator contributions to predicted treatment outcomes are additive, and the current evidence provides no rationale for prioritizing moderators within an additive combination. In other words, there is no empirical support for claims that a gain in phonics knowledge is a necessary-but-not-sufficient component of learning to read in all instructional ecologies. Yet Stuebing et al. (2008) make that assumption: “larger effect sizes were associated with systematic phonics, regardless of the levels of systematic literacy activities and tutoring” (p. 132). One could as easily say that larger effect sizes were associated with systematic literacy instruction, but this statement is not accurate either. In fact, meta-analyses are as useful for identifying and clarifying important research hypotheses and designing experiments to challenge those hypotheses as they are for summarizing the effects of empirical investigations.

One school of thought in reading research is consistent with the notion that phonics instruction is the predominant and necessary (but not sufficient) component of effective reading instruction. Stuebing et al. maintain a slightly different version of this message:

When phonics is systematic (as defined by the NRP), additional well-conceived literacy activities (as defined by Camilli et al, 2003, 2006) are added, and tutoring is used to increase intensity, the effects may be larger than for any of these components in isolation. (p 133)

Doubtlessly, this is true for some reading interventions. Unfortunately, the NRP database does not permit interpretations that support the ordering implied in the above quotation. This is not to say that the evidence refutes this assertion, but rather that studies often did not describe control groups well enough to determine what kinds of instruction were being provided to children, much less to serve as a basis for understanding what types of instruction were contingent on other types of instruction.

Differences aside, we turn to a substantial point of agreement. One important conclusion of Stuebing et al. was that

pedagogical principles exist on a continuum, and should not be dichotomized. In examining this continuum for instruction involving the alphabetic principle, it may be more that the more important component is explicitness and the deliberate attempt to instruct the child as opposed to a scripted approach to phonics (p. 132).

We think this insightful observation is consistent with the current research literature. As noted by Stuebing et al., it is important that those who implement instructional policy do not interpret the word systematic as “scripted.” It is also important to understand that the “explicit” instruction need not be limited to phonics. As aptly put by Stuebing et al. (p. 133) “Creating a scope and sequence, using decodable text, and other ways of systematizing instruction make instruction explicit, but explicitness can be achieved in other ways.”

In contrast to this balanced statement, we believe that some conclusions the NRP extended well beyond the limitations of the data, and we repeat the quotation from above (NRP, 2000b) for emphasis. The NRP report misled readers when it claimed as follows:

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. (p. 2-132)

First, what the NRP called reading was not reading as most of the public thinks of it, but rather it is better described as calling out words in isolation (Camilli & Wolfe, 2004). This oft-repeated quotation above has led to much public confusion (Camilli, Wolfe & Smith, 2006; Garan, 2001). Second, the average advantage of systematic phonics instruction is small relative to other kinds of educational interventions, however poorly described, that emphasize alternatives to phonics instruction. This result should not be shocking. After all, these alternative interventions were devised by researchers to increase reading achievement, not to harm children. Moreover, it is likely that phonics instruction did occur—whether reported or not—and was combined with other beneficial reading activities. In our opinion, the currently available data do not provide an adequate basis for claiming the impact of phonics instruction rises above the panoply of other instructional activities that occur in classrooms. In this regard, an important problem that needs to be addressed is that quantitative studies to date have rarely been successful in providing sufficient detail of treatment implementation in reading research. There is an urgent need for future research to employ appropriate methods that can effectively document this critical information.

Discussion

Critics of our work have previously made much of the notion that the database in *Teaching Children to Read* was not designed to answer questions about the efficacy of language and tutoring (e.g., Francis, 2003, as cited in Archibald, 2003). Along these lines, Stuebing et al. (2008) claimed that “the literature search was not designed to address the comparisons made in Camilli et al. (2003)” (p. 124). This claim would also appear to be the main justification for the NRP’s implicit addition in which an exogenous intercept to the systematic phonics effect. The reasoning seems to be that because the literature search was intentionally designed to measure phonics effects only, then all other effects are subordinate to the phonics interventions. The sum of effects is then claimed to arise from phonics instruction.

This is a relatively weak argument for three reasons. First, the intentions of those who design a database do not retroactively determine the information available in the studies comprising that database. In fact, it is remarkable that incidental moderators (according to the above logic) have larger effects than the central independent variable. It is also quite clear that in some studies in which systematic phonics instruction occurred, such “incidental” variables constituted the central intervention in the instructional ecology (e.g., Tunmer & Hoover, 1993). Second, there are serious design errors in the database (NRP, 2000) that would preclude the determination that it was carefully constructed for any purpose. For example, the study by Vickery, Reynolds, and Cochran (1987) is a pretest-posttest study which did not meet the NRP inclusion criteria (experimental or quasi-experimental design), yet this study provided 8 of 66 contrasts in the NRP database. Moreover, three studies identified by the NRP that met documented inclusion criteria were inexplicably omitted (Camilli et al., 2003). It is unclear how such serious errors survived the peer review process in place at the National Institute of Child Health and Human Development. Shanahan (n.d.) wrote

The panel has agreed upon and shared with Congress an explicit methodological plan that describes virtually every aspect of the study underway. If successfully accomplished, this plan will represent one of the most thorough, careful, and rigorous analyses of reading data ever conducted. (web edition)

In retrospect, it does not appear that this intent was successfully accomplished, and to our knowledge, there has been no effort to correct obvious inconsistencies in the original NRP database.

Another argument given by Stuebing et al. is that because the study was designed to collect information on phonics treatments only, the sample of treatments for tutoring and language is “not a random sample from the population” (p. 130). A simple thought experiment would suggest that if an researcher designed a study focused on phonics interventions, then the strongest effect should be found for those interventions. Thus, the nonrepresentative sample of studies for language and tutoring could easily be construed to result in underestimating their effect sizes. In summary, we think there is little empirical or logical support for the argument that the database design—or more accurately the intentions of those who designed the database—should be privileged as a framework for interpretations of statistical results.

Systematic phonics instruction may yet prove in a public fashion to be the most essential component of reading instruction, yet the current evidence is unconvincing on this point. This result may seem incomprehensible to many knowledgeable researchers who believe that explicit instruction in phonics is a necessary and effective intervention for some early readers. Yet it is important to understand that the NRP’s main finding about the superiority of phonics instruction hinged on a semantic construction that was unsuccessful in compensating for incomplete descriptions of treatment and control group interventions in the research literature. Likewise, the attempt to resolve

incomplete reporting by assuming that treatment moderators were nested in (or contingent on) phonics instruction was also unsuccessful. The current literature does not provide a strong indication of how instructional activities should be prioritized and combined, though good teachers know how to do this given appropriate training in a range of instructional strategies (Camilli & Wolfe, 2004).

The issues that we have raised require careful thinking through by those who would choose to be persuaded by evidence in designing school-based instruction. To be clear, we found that descriptive information was underreported for alternatives to phonics interventions, and on this basis, we contend that the truth of the NRP's claim with regard to programmatic instruction cannot be conclusively established. The claim relative to standard instruction is stronger (to the degree that standard instruction provides an appropriate comparison point), but there is less of a case for generalizing to the diverse contexts in which "standard instruction" occurs. Based on our analysis of the NRP database, we do not claim that future programs that do or do not emphasize phonics will have equivalent outcomes. We also do not make general claims about the relative benefits of programs with phonics or literacy emphases. Meta-analysis of the current database permits smaller-scale conclusions than those advanced in the various commentaries on our work. Accordingly, there is evidence that suggests literacy activities can have a beneficial effect on measured reading outcomes, as can phonics activities and tutoring. In any case, the new study by Stuebing et al. (2008) has helped to sharpen our thinking regarding the original findings of Camilli et al. (2003), and for this reason, we believe it has contributed to a deeper understanding of the research literature on learning to read.

References

- Archibald, G. (2003, June 10). Researchers verify reading ability gets a boost from phonics. *Washington Times*. Retrieved June 15, 2003, from <http://www.washingtontimes.com/national/20030610-125708-5569r.htm>.
- Baumann, J. F., Hoffman, J. V., Moon, J. S., & Duffy-Hester, A. (1998). Where are the teachers' voices in the phonics/whole language debate? Results from a survey of U.S. elementary classroom teachers. *Reading Teacher*, 51(8), 636–650.
- Briggs, D. (2007). Synthesizing casual inferences. *Educational Researcher*, 37, 15–22.
- Camilli, G., Vargas, S., & Yurecko, M. (2003). Teaching children to read: The fragile link between science and federal education policy. *Education Policy Analysis Archive*, 11(15). Retrieved March 20, 2007, from <http://epaa.asu.edu/epaa/v11n15/>.
- Camilli, G., & Wolfe, P.M. (2004). Meta-analysis and instructional policies in reading. *Educational Leadership*, 61(6), 26–29.
- Camilli, G., Wolfe, P.M., & Smith, M.L. (2006). Meta-analysis and reading policy: Perspectives on teaching children to read. *The Elementary School Journal*, 107, 27–36.
- Cook, T. (1997). Lessons learned in evaluation over the last 25 years. In E. Chelimsky & W.R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 30–52). Thousand Oaks, CA: SAGE.
- Cooper, H. G. (2005). Reading between the lines: Observations on the report of the National Reading Panel and its critics. *Phi Delta Kappan*, 86, 456–461.
- Dawson, R., & Tilley, N. (1997). An introduction to scientific realist evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 405–418). Thousand Oaks, CA: SAGE.
- Fisher, D., Lapp, D., & Flood, J. (1999). How is phonics really taught? *Yearbook of the National Reading Conference*, 48, 134–145.
- 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.
- Garan, E.M. (2001). Beyond the Smoke and Mirrors. *Phi Delta Kappan*, 82, 500–506.
- Hammill, D. D., & Swanson, H. L. (2006). The National Reading Panel's meta-analysis of phonics instruction: Another point of view. *Elementary School Journal*, 107, 17–26.
- Morrow, L. M., & Tracey, D. H. (1997). Strategies used for phonics instruction in early childhood classrooms. *Reading Teacher*, 50(8), 644–651.

- National Institute of Child Health and Human Development (2000). *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* (NIH Publication No. 00-4769). Washington, DC: U. S. Government Printing Office.
- Pressley, M., Rankin, J., & Yokoi, L. (1996). A survey of instructional practices of primary teachers nominated as effective in promoting literacy. *Elementary School Journal, 96*(4), 363–384.
- Scriven, M. (1981). Product evaluation. In N. L. Smith (Ed.), *New Techniques for evaluation* (pp. 121–166). Beverly Hills, CA: Sage.
- Shanahan, T. (n. d.). The national reading panel: Using research to create more literate students. *Reading Online*. Retrieved November 20, 2007, from <http://www.readingonline.org/critical/shanahan/panel.html>.
- Stuebing, K. K., Barth, A. E., Cirino, P. T., Francis, D. J., & Fletcher, J. M. (2008). A response to recent re-analyses of the National Reading Panel Report: Effects of systematic phonics instruction are practically significant. *Journal of Educational Psychology, 100*, 123–134.
- Torgesen, J., Alexander, A., Wagner, R. Rashotte, C. Voeller, K., & Conway, T. (2001). Intensive Remedial instruction for children with reading disabilities. *Journal of Learning Disabilities, 34*, 32–58.
- 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.

About the Authors

Gregory Camilli

Rutgers, the State University of New Jersey

Sun Hee Kim

Teachers College, Columbia University

Sadako Vargas

Rutgers, the State University of New Jersey

E-mail: camilli@rci.rutgers.edu

Gregory Camilli is Professor of Educational Statistics and Measurement in the Graduate School of Education at Rutgers, the State University of New Jersey. His current research interests include meta-analysis, educational effectiveness, differential item functioning, and affirmative action in law school admission. His publications include *Summarizing Item Difficulty Variation with Parcel Scores* (Camilli, G., Prowker, A.N., Dossey, J., Lindquist, M., Chiu, T.W., Vargas, S. & de la Torre, forthcoming); *Illustration of a multilevel model for meta-analysis* (de la Torre, J., Camilli, G., Vargas, S., & Vernon, R. F., 2007); and *Handbook of Complementary Methods in Education Research* (Green, J. L., Camilli, G., & Elmore, P.B., 2006).

Sun Hee Kim is a doctorate student in Measurement and Evaluation in the Department of Human Development, Teachers College, Columbia University. Her research focuses on measurement and statistical analysis in psychology, education, and policy. Her publications include examining individual change over time in an alternative assessment, and relationships between literacy and language ability of young children.

Sadako Vargas is Research Associate in the Graduate School of Education at Rutgers, the State University of New Jersey. Her research interests meta-analysis and the effectiveness of occupational therapy. Her publications include *A meta-analysis of research on sensory integration therapy* (Vargas, S. & Camilli, 1999); *The origin of the National Reading Panel: A response to "Effects of systematic phonics instruction are practically significant"* (Camilli, G., Kim S.H. & Vargas, S., forthcoming); and *Teaching Children to Read: The fragile link between science and federal education policy* (Camilli, G., Vargas, S., and Yurecko, M., 2003).

EDUCATION POLICY ANALYSIS ARCHIVES <http://epaa.asu.edu>

Editor: Sherman Dorn, University of South Florida

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Sherman Dorn, epaa-editor@shermamdorn.com.

Editorial Board

| | |
|------------------------------|-------------------------------|
| Noga Admon | Jessica Allen |
| Cheryl Aman | Michael W. Apple |
| David C. Berliner | Damian Betebenner |
| Robert Bickel | Robert Bifulco |
| Anne Black | Henry Braun |
| Nick Burbules | Marisa Cannata |
| Casey Cobb | Arnold Danzig |
| Linda Darling-Hammond | Chad d'Entremont |
| John Diamond | Amy Garrett Dikkers |
| Tara Donohue | Gunapala Edirisooriya |
| Camille Farrington | Gustavo Fischman |
| Chris Frey | Richard Garlikov |
| Misty Ginicola | Gene V Glass |
| Harvey Goldstein | Jake Gross |
| Hee Kyung Hong | Aimee Howley |
| Craig B. Howley | William Hunter |
| Jaekyung Lee | Benjamin Levin |
| Jennifer Lloyd | Sarah Lubienski |
| Les McLean | Roslyn Arlin Mickelson |
| Heinrich Mintrop | Shereeza Mohammed |
| Michele Moses | Sharon L. Nichols |
| Sean Reardon | A.G. Rud |
| Ben Superfine | Cally Waite |
| John Weathers | Kevin Welner |
| Ed Wiley | Terrence G. Wiley |
| Kyo Yamashiro | Stuart Yeh |

EDUCATION POLICY ANALYSIS ARCHIVES <http://epaa.asu.edu>

**New Scholar Board
English Language Articles
2007–2009**

| | |
|---------------------------|------------------------|
| Wendy Chi | Corinna Crane |
| Jenny DeMonte | Craig Esposito |
| Timothy Ford | Samara Foster |
| Melissa L. Freeman | Kimberly Howard |
| Nils Kauffman | Felicia Sanders |
| Kenzo Sung | Tina Trujillo |
| Larisa Warhol | |

Archivos Analíticos de Políticas Educativas <http://epaa.asu.edu>

Editor

Gustavo E. Fischman

Arizona State University

Asistentes editoriales: Rafael O. Serrano (ASU) & Lucia Terra (UBC)

Hugo Aboites

UAM-Xochimilco, México

Claudio Almonacid Avila

UMCE, Chile

Alejandra Birgin

FLACSO-UBA, Argentina

Mariano Fernández Enguita

Universidad de Salamanca. España

Roberto Leher

UFRJ, Brasil

Pia Lindquist Wong

CSUS, USA

Alma Maldonado

University of Arizona, USA

Imanol Ordorika

IIE-UNAM, México

Miguel A. Pereyra

Universidad de Granada, España

Romualdo Portella de Oliveira

Universidade de São Paulo, Brasil

José Ignacio Rivas Flores

Universidad de Málaga, España

José Gimeno Sacristán

Universidad de Valencia, España

Susan Street

CIESAS Occidente, México

Daniel Suárez

LPP-UBA, Argentina

Jurjo Torres Santomé

Universidad de la Coruña, España

Armando Alcántara Santuario

CESU, México

Dalila Andrade de Oliveira

UFMG, Brasil

Sigfredo Chiroque

IPP, Perú

Gaudêncio Frigotto

UERJ, Brasil

Nilma Lino Gomes

UFMG, Brasil

María Loreto Egaña

PIIE, Chile

José Felipe Martínez Fernández

UCLA, USA

Vanilda Paiva

UERJ, Brasil

Mónica Pini

UNSAM, Argentina

Paula Razquin

UNESCO, Francia

Diana Rhoten

SSRC, USA

Daniel Schugurensky

UT-OISE Canadá

Nelly P. Stromquist

USC, USA

Antonio Teodoro

Universidade Lusófona, Lisboa

Lílian do Valle

UERJ, Brasil