Journal of Educational Psychology
2008, Vol. 100, No. 1, 123–134

0022-0663/08/$12.00 DOI: 10.1037/0022-0663.100.1.123

A Response to Recent Reanalyses of the National Reading Panel Report: Effects of Systematic Phonics Instruction Are Practically Significant

Karla K. Stuebing, Amy E. Barth, Paul T. Cirino, David J. Francis, and Jack M. Fletcher
University of Houston

The authors examine the reassessments of the National Reading Panel (NRP) report (National Institute of Child Health and Human Development, 2000) by G. Camilli, S. Vargas, and M. Yurecko (2003); G. Camilli, P. M. Wolfe, and M. L. Smith (2006); and D. D. Hammill and H. L. Swanson (2006) that disagreed with the NRP on the magnitude of the effect of systematic phonics instruction. Using the coding of the NRP studies by Camilli et al. (2003, 2006), multilevel regression analyses show that their findings do not contradict the NRP findings of effect sizes in the small to moderate range favoring systematic phonics. Extending Camilli et al. (2003, 2006), the largest effects are associated with reading instruction enhanced with components that increase comprehensiveness and intensity. In contrast to Hammill and Swanson, binomial effect size displays show that effect sizes of the magnitude found for systematic phonics by the NRP are meaningful and could result in significant improvement for many students depending on the base rate of struggling readers and the size of the effect. Camilli et al. (2003, 2006) and Hammill and Swanson do not contradict the NRP report, concurring in supporting comprehensive approaches to reading instruction.

Keywords: reading instruction, phonics, National Reading Panel, meta-analysis

The report of the National Reading Panel (NRP; National Institute of Child Health and Human Development [NICHD], 2000), a congressionally mandated effort to synthesize research on effective instructional methods for teaching children to read, continues to generate controversy. The NRP report was completed by a committee under the direction of the NICHD in collaboration with the U.S. Department of Education. Despite the use of an empirical approach to the synthesis of research results (meta-analysis), with two peer-reviewed articles representing the syntheses of phonics and phonological awareness instruction (Ehri, Nunes, Stahl, & Willows, 2001; Ehri, Nunes, Willows, et al., 2001), the report, and especially the part involving phonics, has been highly scrutinized since it was released in 2000. Several reviewers have disagreed with NRP conclusions supporting the efficacy of systematic phonics instruction over other approaches to teaching phonics (Allington, 2002; Garan, 2001; see responses by Cooper, 2005; Shanahan, 2004).

Recently, two critiques of the phonics meta-analysis of the NRP report were published in a special issue of the *Elementary School Journal* (Camilli, Wolfe, & Smith, 2006; Hammill & Swanson, 2006). Camilli et al. (2006) provided a second reanalysis of the studies coded by the NRP. In their first reanalysis of the NRP report, Camilli, Vargas, and Yurecko (2003) concluded that the actual effect size from studies involving systematic phonics instruction in the NRP report was $d = 0.24$ when studies were weighted equally (i.e., without correction for sample size) and $d = 0.188$ when studies were weighted by a combination of equal representation and sample size. These estimates were much lower than the $d = 0.41$ reported by the NRP for end-of-training outcomes. In Camilli et al. (2006), more extensive coding of study characteristics was provided, along with appropriate multilevel analysis techniques, leading to $d = 0.123$ for systematic phonics, which was characterized as not significant and a "weak intervention" (p. 31).

In a similar vein, Hammill and Swanson (2006) converted the estimates of effect size in the NRP report to metrics (R^2) that represent the amount of explained variance. Although acknowledging that "94% of the *d*'s supported the superiority of phonics instruction over other approaches," they went on to observe that "Cohen would describe 65% of these significant *d*'s as small" (p. 19). Converting the *d*s to *r*s yielded an overall $r = .21$, or R^2 of .04, "suggesting that 96% of the variance in reading achievement can be attributed to factors other than the systematic phonics instruction" (p. 18). The authors concluded that "for all practical purposes, the advantages of phonics versus nonphonics instruction have not been demonstrated" (p. 25).

Theoretical and Pedagogical Issues

Underlying the controversy over phonics and its role in reading instruction is a set of theoretical issues about learning to read that relate directly to how the alphabetic principle is taught (American

Karla K. Stuebing, Amy E. Barth, Paul T. Cirino, and David J. Francis, Department of Psychology; Texas Institute for Measurement, Evaluation and Statistics; and the Texas Center for Learning Disabilities, University of Houston; Jack M. Fletcher, Department of Psychology and the Texas Center for Learning Disabilities, University of Houston.

Grant P50 HD052117 from the National Institute of Child Health and Human Development to the Texas Center for Learning Disabilities supported this article. David J. Francis was a methodological consultant to the National Reading Panel (NRP). No other authors were involved with the NRP and its report.

Correspondence concerning this article should be addressed to Jack M. Fletcher, Department of Psychology, University of Houston TMC Annex, 2151 West Holcombe Boulevard, Suite 222, Houston, TX 77204-5355. E-mail: jackfletcher@uh.edu

Federation of Teachers, 1999; Pressley, 2005; Stanovich, 2000). Unlike in previous periods, the current discussion is rarely about whether any instruction involving the alphabetic principle should be provided, but (a) how systematically instruction should be conducted to ensure that all students have adequate knowledge of the alphabetic principle, and (b) the extent to which students are better served by opportunities to make inferences and develop their own understanding of the role of the alphabetic principle versus opportunities in which the alphabetic principle is directly taught by the teacher (Allington, 2002; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). These differences are often represented as a dichotomy comparing a scripted curriculum with a defined scope and sequence versus a curriculum that encourages discovery of the alphabetic principle through immersion in literature.

Underlying this dichotomy is a continuum that reflects the extent to which the student is expected to infer and construct new knowledge versus learn through the direct, explicit sharing of the new knowledge by the teacher. The NRP coded studies involving phonics to capture this continuum, comparing studies that included "systematic phonics," characterized by "a planned, sequential introduction of a set of phonics elements along with teaching and practice of these elements" (NICHD, 2000, p. 2-89), with those that included "unsystematic phonics," in which there was evidence of some phonics instruction but not in a planned, sequential fashion, and "no phonics," in which there was no evidence of any attempt to teach phonics except incidentally. In the latter programs, the instruction might focus on the teaching of whole words or might incidentally address the development of phonics skills within reading, writing, listening, and speaking activities as the need arises (NICHD, 2000). The overall effect size reported by the NRP favored systematic phonics over other forms of phonics instruction.

Estimation of the NRP Effect Sizes: Hypothesis 1

The studies by Camilli et al. (2003, 2006) and Hammill and Swanson (2006) are empirical reassessments of the NRP results. Both essentially minimize the findings of the NRP on the efficacy of reading instruction that includes systematic phonics and therefore downplay instructional theory and pedagogy that focuses on planned, systematic instruction in the alphabetic principle for all students.

If the level of systematicity of phonics instruction is primarily responsible for the treatment effect, a dosage hypothesis may be in effect. The largest effect sizes should be associated with the strongest dose – or systematic phonics compared to the no phonics control. Smaller effect sizes should be associated with comparisons of systematic phonics with unsystematic phonics and also with comparisons of unsystematic phonics with no phonics. This simple main effects hypothesis was the initial comparison made in the NRP report and as an average, does not take into account the possibility that the instruction might interact with child characteristics; some children who are weaker in alphabetic skills may need more explicit phonics instruction (e.g., Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998), whereas the degree of systematic phonics instruction may be less important for other children with better developed letter–sound knowledge. This main effect hypothesis also ignores the fact that type of instruction may interact with school and teacher characteristics (e.g., teaching competence, time allocation; Connor et al., 2007; Foorman et al., 2006). Nonetheless, the dosage hypothesis is a useful heuristic for understanding the NRP report and its subsequent evaluations.

In calculating effect sizes, the NRP held the treatment constant, comparing the systematic phonics group from each study to any available comparison group. Some of the effect sizes represent contrasts between systematic phonics and some phonics, whereas other effect sizes represent contrasts between systematic phonics and true no phonics control groups. The d of 0.41 is the average of two groups of effects: (a) systematic phonics contrasted with some phonics, and (c) systematic phonics contrasted with no phonics. The NRP report did break down effects using type of comparison group as a moderator. The largest effects represented comparisons of systematic phonics to whole-word instruction (average $d = 0.51$), and the smallest average effects represented comparisons wherein the control group had "whole language" instruction ($d = 0.31$).

Camilli et al. (2003) did not accept the NRP premise that the average of different levels of phonics intervention was a meaningful comparison, focusing on study characteristics that might moderate the NRP effect size estimates. They identified two effect sizes that they felt had been miscalculated and disagreed with the inclusion of one study (Vickery, Reynolds, & Cochran, 1987) that they felt did not meet inclusion criteria and was noteworthy because it contributed eight effect sizes to the NRP database. They also identified studies in the NRP database that they felt should have been included by the NRP. Based on these modifications, Camilli et al. (2003) then recoded all of the studies from the set used by the NRP to make a different set of comparisons. They held constant the control group—always using the no phonics control and contrasting it to what they coded as "some phonics" treatments and "systematic phonics" treatments. Additionally, they coded comparisons across all groups in each study. For example, if there was a group that received a whole-word treatment, it would be compared to the no treatment (or standard practice) control group. As a result, Camilli et al. (2003) produced a much larger number of effects from these studies by coding for comparisons among all group means, which included some comparisons for which there was no phonics in either the treatment or control group. They also generated different estimates of effect size because the comparisons were adjusted to account for the presence of study moderators.

A potential limitation of these estimates is that the literature search was not designed to address the comparisons made in Camilli et al. (2003). Studies that compared no phonics and some phonics would not necessarily have been identified by the NRP search strategy. Additionally, the NRP search strategy was not designed to locate studies that compared no phonics or some phonics to a standard treatment control with systematic language activities. Finally, the NRP estimated the effect of systematic phonics instruction by taking weighted averages of the effects, both overall and within specific hypothesized moderator groups. Camilli et al. (2003) estimated the effects of systematic phonics instruction by predicting the effects via a regression equation that included recoding of the NRP studies for the amount and degree of systematicity of phonics in the treatment; whether there were "systematic language activities" in the treatment group and/or control group; and the intensity of the treatment delivery, repre-

sented as tutoring versus small-group-/classroom-level delivery. The effect size compared to the NRP overall $d = 0.41$ was a regression weight that represented the difference between systematic phonics and some phonics controlling for the other intervention characteristics that were in the model.

Altogether, the NRP report and Camilli et al. (2003) ask different questions. The NRP question is analogous to asking about the value of receiving the intervention versus not receiving the intervention. The Camilli et al. (2003) report is analogous to asking what is the value of receiving a strong form of the intervention compared to receiving weaker forms of the intervention and relative to factors that moderate the outcomes. From our view, both questions are reasonable for intervention studies; it is worthwhile to know the average effect of supplying the intervention (i.e., the NRP report approach), but it is also worthwhile to know the incremental value of adding the intervention over and above the value of other characteristics of students, classrooms, and teachers that might affect outcomes (i.e., the Camilli et al. [2003] approach).

One would expect the estimates of effect size from Camilli et al. (2003) and the NRP to differ because the questions were different and (a) the results were based on a slightly different set of studies; (b) a different approach was used to code effects; and (c) the overall NRP effect was an unadjusted effect, whereas the Camilli et al. (2003, 2006) effects were adjusted for other study characteristics. Our first hypothesis is that when the same question is asked of the Camilli et al. (2003) data (i.e., the parameters estimated are the same), the magnitude of the effects of systematic phonics in Camilli et al. (2003, 2006) will be comparable to that of the NRP report.

Comprehensive Approaches to Reading Instruction: Hypothesis 2

In addition to the NRP report, other consensus reports, including the National Research Council report (Snow, Burns, & Griffin, 1998) and the report of the RAND Reading Study Group (2002), have argued for comprehensive approaches to reading instruction that include literacy activities that extend beyond alphabetics. This question was tested by Camilli et al. (2003) and further addressed by Camilli et al. (2006), who employed a multilevel analysis to account for the dependence of multiple effect sizes from the same studies. As discussed previously, Camilli et al. (2003) also coded for other components of instruction that were wrapped in with phonics instruction in the treatment, including systematic language activities and intensity. Note that the coding of language activities in Camilli et al. (2006) used "systematic" in a different manner from the NRP, referring primarily to the presence of literacy activities believed to facilitate fluency and comprehension and not to the degree of explicitness. These activities represented the extent to which teachers incorporated components of reading that extend beyond alphabetics, emphasizing a print-rich environment, independent reading, purposeful writing, the use of invented spelling, and the use of literature to teach higher order skills. Tutoring is not really another component of reading but is a means of delivering instruction in a manner that allows for more practice at a smaller teacher–student ratio, thus increasing the intensity of instruction.

In the final models presented by Camilli et al. (2003, 2006), the individual components of the instruction all added significantly to the prediction of the effect size. Not emphasized was the implication that the overall effect size for a more comprehensive approach could be quite large. To address this question we repeated the multilevel regression analysis with modifications to the data set that permitted a more straightforward presentation of the value of adding additional activities to the reading program. Our second hypothesis is that in this stricter comparison, which allows separate empirically derived estimates of the effect of additional literacy activities on top of both some phonics and systematic phonics, effect sizes will be larger when additional literacy activities and tutoring are added to the effects of systematic phonics instruction.

Effect Sizes in Context: Hypothesis 3

Even the critics of the NRP report concede that different studies show small effects significantly favoring systematic phonics; they disagree that these effect sizes are practically important, which is the point of departure for Hammill and Swanson (2006). Evaluating the practical utility of treatment effects is an important issue, which Hammill and Swanson addressed by translating the *d*s into their associated *r*s and R^2s and then classifying these values using an idiosyncratic variation on Cohen's suggested rules of thumb for choosing an effect size for a priori estimation of statistical power. We do not question the general heuristic labeling of effect sizes as *small* ($r = .1$ and $d = 0.2$), *medium* ($r = .3$ and $d = 0.5$), or *large* ($r = .5$ and $d = 0.8$) (Cohen, 1988) but question the use of these heuristics by Hammill and Swanson because they were never intended to be strict rules or thresholds for judging the utility of intervention effect sizes.

Cohen proposed that when conducting a power analysis in the planning stages of research, the researcher should first look to prior research in the same field to get an idea of the size of an effect that might be expected, which permits an estimate of the sample size that will be required to minimize the risks of a Type II error (in this context, identifying a treatment as ineffective when in fact it was effective). If the researcher is not able to obtain a numeric value to use in the power analysis calculations, Cohen (1988) proposed the use of the now familiar effect size heuristics but warned against their misuse and abuse. He stated:

> The author proposes *as a convention*, [effect size] values to serve as operational definitions of the qualitative adjectives "small", "medium", and "large." This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as "large" are sometimes understood as absolute, sometimes as relative; and thus, they run the risk of being misunderstood. (p. 12)

Cohen (1988) also insisted that the importance of an effect could not be discerned from its size independently of its context, indicating that "the *meaning* of any given [effect size] is, in the final analysis, a function of the context in which it is embedded" (p. 535). Small effect sizes for important outcomes may have significant implications in a practical context (Trusty, Thompson, & Petrocelli, 2004). Furthermore, small effects in ongoing processes, such as the development of key skills underlying reading, may accrue over time to become moderate to large effects (Prentice & Miller, 1992). Conversely, large effect sizes may be trivial within practical contexts because the spuriously large effect was the result

of method variance and response biases (Thorndike, 1997) and from specification error and outliers (Pedhazur, 1982). As Cohen (p. 535) stated,

> "only 50% of the variance" may be as valid a formulation in one context as "only 1% of the variance" is in another, and conversely, "as much as 1% of the variance" is, in principle, no less valid a formulation than "as much as 50% of the variance."

Thus, one-size-fits-all rules of thumb may not help to interpret an effect size in the absence of the context in which it is to be operationalized (Vacha-Haase & Thompson, 2004).

Considering the caveats provided by Cohen (1988), the labeling of effect sizes can serve as an aid to communication, in the sense that a large effect from a high-quality study is nearly always "better" than a small or medium effect. Although difficulty in making decisions about the practical impact of a given effect size is less likely in the case of an effect size with a large value (e.g., in the range of about $d = 0.80$), even here one might envision a hypothetical situation in which, given a significant cost to obtain the result, an effect size of this absolute magnitude might carry little practical meaning. For example, individual tutoring may have a larger effect size than classroom-based instruction, but providing 1:1 instruction to every single student in public schools carries an unrealistic cost, not to mention the logistical difficulties of training sufficient numbers of teachers, rebuilding schools to support tutoring, and so on. However, as effect sizes become progressively lower in absolute value, an evaluation of contextual factors is more likely to be required, which is more complicated than simply comparing an effect size to some benchmark. A consideration of costs and benefits may be required. Although labels can be used with effect sizes, the utility of the effect under consideration is not synonymous with its label. For example, although the translation of *small* into $d = 0.2$ could be useful when planning research, the translation of *small* into *of little utility and of little practical importance* is a much more tenuous proposition and is likely to be biased without a consideration of context.

A clear example of such an effect is the well-known 5-year randomized study that led to the recommendation to utilize aspirin in the prevention of heart attack (Steering Committee of the Physicians' Health Study Research Group, 1989). The absolute value of the effect size from this study was $d = 0.07$, which is *negligible* under heuristic guidelines of standard deviation units and certainly unimpressive in terms of correlation ($r = .034$) or in terms of proportion of variance ($R^2 = .001$). However, given the low base rate of heart attack (1.33% in that study), the effect size of $d = 0.07$ means that nearly twice as many individuals in the placebo group suffered a heart attack relative to individuals in the aspirin condition. In the context of the low base rate and high costs of heart attacks and the generally low cost of aspirin, this effect size moves from unimpressive to powerful.

The average effect sizes of different educational interventions are larger than that in the aspirin example. Lipsey and Wilson (1993) reported that the average educational intervention had an effect size of $d = 0.34$, or $r = .168$, which is R^2 of .028. In another large meta-analysis, Swanson, Hoskyn, and Lee (1999) reported an average weighted effect size of $d = 0.56$ for the effects of educational interventions in people identified with learning disabilities across a wide variety of outcome measures. The latter effect size is larger than the average effect for phonics in the NRP report, which in turn is larger than that of Lipsey and Wilson. Potential explanations for the variability of these results include differences in the interventions; the populations, which differed by age and subgroup; as well as methodological quality; research design; and so on. The Lipsey and Wilson meta-analysis involved a variety of educational interventions in kindergarten through Grade 12 and college. In contrast, the effect size reported by Swanson et al. was for individuals with learning disabilities in kindergarten through Grade 12. The NRP effect size of $d = 0.41$ was for phonics interventions in kindergarten through Grade 6 over all subgroups. However, note that for poor readers, the NRP reported an effect size of $d = 0.98$ for decoding regular words and $d = 0.67$ for decoding pseudowords in kindergarten through Grade 2 and $d = 0.49$ for decoding regular words and $d = 0.52$ for decoding pseudowords in Grades 2 through 6. These latter values are especially comparable to the $d = 0.57$ reported by Swanson et al. for the effect of intervention in groups with learning disabilities on word recognition. Obviously none of these meta-analyses were exact replications of one another. The amount of variation in these estimates is consistent with the possible effects of the moderators mentioned previously as well as expected sampling error. The next step, however, is not to compare the obtained effect to a one-size-fits-all rule of thumb or convention, but to use these estimated effects to determine the utility of implementing various types of interventions by taking into account contextual factors.

In the case of interventions designed to improve reading performance, identifying factors such as the base rate of struggling readers and the cost of interventions as well as the cost of not providing an intervention (e.g., the cost that might be represented by dropping out of high school) can assist in contextualizing the effect sizes found by either the NRP study or the Camilli et al. (2003, 2006) analyses. To address this question, we used the effect size estimates from the NRP report and different estimates of the incidence of reading difficulties and a variant of a binomial effect size display (Rosenthal & Rubin, 1982) to evaluate the importance of context in interpreting effect size data. Consistent with Cohen (1988), our third hypothesis is that interventions with effect sizes as small as those identified by Hammill and Swanson (2006) for phonics instruction could significantly reduce the number of children with reading problems depending on the base rate used to estimate the incidence of reading difficulties and the effect size associated with different interventions.

Hypothesis 1

Method

Database. To address the first hypothesis concerning the impact of study parameters on the discrepancies in the NRP report and in Camilli et al. (2003), we relied on the corpus of studies identified by the NRP and recoded by Camilli et al. (2003, 2006). These studies and the database of effect sizes and codes were accessed from the online journal *Education Policy Archives Analysis* Web site in which Camilli et al. (2003) originally appeared (http://epaa.asu.edu/epaa/v11n15). Although several questions were asked in the section of the NRP report addressing phonics, the first question was the target of both Camilli et al. (2006) and Hammill and Swanson (2006): "Does systematic phonics instruction help children learn to read more effectively than nonsystem-

atic phonics instruction or instruction teaching no phonics?" (NICHD, 2000, p. 2-92). The primary criterion for including studies into this meta-analysis was stated in the NRP methodology: "Studies had to provide data testing the hypothesis that systematic phonics instruction improves reading performance more than instruction providing unsystematic phonics or no phonics instruction." (NICHD, 2000, p. 2-90). This criterion requires that a study include a comparison of systematic phonics instruction versus any other control condition, which is what the $d = 0.41$ in the NRP report represents. Some of these control conditions will have no phonics and others will have unsystematic phonics. With this strategy and other criteria, the NRP screened 75 studies representing randomized or quasi-experimental designs with treatment and control groups, with 38 represented in the final database. Camilli et al. (2003) included substantially the same set of studies included in the NRP meta-analysis, removing one study and adding three from the corpus of studies identified by the NRP (no new search was conducted).

Predictors. The dependent variable in Camilli et al. (2003) was the effect size, and the independent variables were coded vectors that represented many other characteristics of the intervention given in each study. After a stepwise regression procedure, their final model included as predictors the amount and degree of systematicity of phonics in the treatment; the presence of systematic language activities in the treatment group and/or control group; and the intensity of the treatment delivery, represented as tutoring versus small-group-/ classroom-level delivery.

Procedures. We reformulated the results from the two studies within the same framework. To further understanding of the dosage hypothesis and what it means for comparing effect size estimates from the NRP report and Camilli et al. (2003), consider Figure 1. The horizontal axis represents performance on some reading outcome measure. We hypothesized that the average performance associated with no phonics would be at the far left (lowest performance), the average performance for systematic phonics would be on the far right (highest performance), and the average performance associated with unsystematic or some phonics would be in between these two extremes. The line segments connecting the three groups represent the distance in effect size units between each pair. Line segment *a* represents the average effect between systematic phonics and a no phonics control group, line segment *b* represents the average effect between some phonics and a no phonics control group, and line segment *c* represents the difference between systematic phonics and some phonics groups. We can use this generic framework to highlight the different elements that are being considered in the NRP report and in the Camilli et al. (2003, 2006) reports and to demonstrate why we would not expect the numbers they report to be the same, even in the context of the same studies.

The NRP asked about the effects of systematic phonics instruction. Therefore, they coded some studies that compared systematic phonics to a no phonics control (represented by line segment *a*) and other studies that compared systematic phonics to a some phonics control (line segment *c*). They then averaged the set of *a* and *c* effects to get $d = 0.41$. Note that the size of this average depends on the average size of *a*, the average size of *c*, and the number of studies providing estimates of *a* and *c* found in the literature. If, consistent with our model, *a*s are systematically larger than *c*s and there are more of them, the overall average will be larger than if the proportion of *a*s and *c*s had been reversed. The NRP also included a test of the homogeneity of the distribution of effects and found that they were not homogenous. When effects are not homogenous, presenting the overall effect is only Step 1, followed by presentation of the effects within moderator groups. As a result, the NRP presented the average effects within many potential moderator groups, including one breakdown that showed that the effect size differed depending on the type of control group.

Camilli et al. (2003) recoded all of the effects to arrive at a different set of comparisons. They made all comparisons against a no phonics control, thus estimating the quantities *a* and *b* in Figure 1. Note that they also coded estimates of the difference between groups where neither of the groups received any phonics instruction. An example would be a comparison between a treatment group receiving whole-word instruction and a control group receiving standard instruction. This contrast cannot be represented in Figure 1, which represents the continuum of the phonics treatment effects. In studies that included a systematic phonics group, a some phonics group, and a no phonics control group, Camilli et al. (2003) coded both difference *a* and difference *b*, thus obtaining more comparisons from the same set of studies than the NRP analysis. They then analyzed the coded effects in a regression

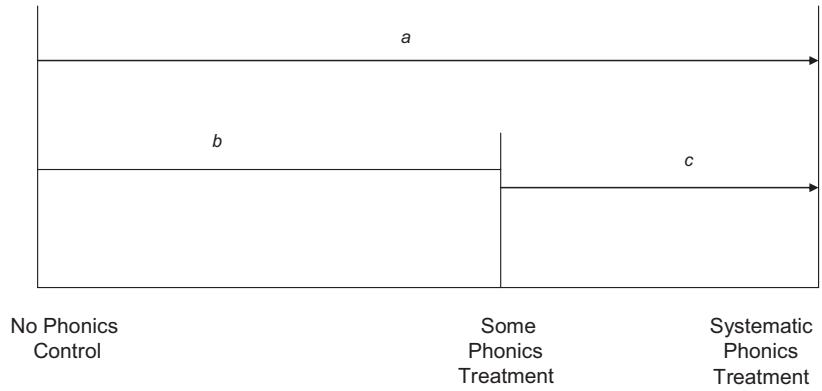

Figure 1. Model for comparisons of effect sizes from the National Reading Panel report and Camilli et al. (2003, 2006).

analysis to estimate and test the *c* difference (i.e., the difference between systematic phonics and some phonics).

The conceptual difference between the two approaches is that the NRP estimated the average of a mixed set of effects (*a* and *c*) and then tested the distribution of these effects to determine if they were homogenous. Camilli et al. (2003) estimated *c*, the average difference between the estimates of effects *a* and *b*, and tested whether there was a significant difference when the set of effects was divided along the some phonics–systematic phonics dimension. The overall average, or the NRP estimate, is almost certainly going to be larger than *c* regardless of the mix of *a*s and *c*s, provided there is at least one *a* effect in the group of estimates. We used the model in Figure 1 as a framework for comparing estimates of effect size in Camilli et al. (2003) and the NRP report.

Results

Camilli et al. (2003) provided several different estimates for *c*. The first was presented in the context of their Table 3, in which the average univariate effect for comparisons of systematic phonics with no phonics controls was $d = 0.514$ and the average comparison of some phonics with no phonics controls was $d = 0.243$. To put this into the context of Figure 1, the former was their estimate of the average *a* effect; the latter was their estimate of the *b* effect. We can calculate the difference between the two to arrive at an estimate of $c = .27$. Camilli et al. (2003) pointed out that their estimate of the effect of systematic phonics was about 30% smaller than the effect reported by the NRP. They failed to underscore the fact that .27 is only an estimate of the average *c* effect and not, like the NRP average, an estimate of the average of the mixed *a* and *c* effects, which Figure 1 shows must be larger. We do not know how many *a* and *c* effects were coded by the NRP because we based our reanalysis on the Camilli et al. (2003) data set, which did not contain this information. However, if we make the simplifying assumption that there are equal numbers of *a* and *c* effects and use these estimates of *a* and *c* (.514 and .27), our estimate of the average effect is (.514 + .27) / 2 or $d = 0.39$, which is remarkably close to the value of $d = 0.41$ obtained in the NRP analysis. Thus, when we estimate the same parameter, the results of the two coding rubrics converge despite slight differences in the corpus of studies.

Other estimates of the *c* effect can be obtained from Camilli et al. (2003). Specifically, their Tables 6 and 7 display the parameter estimates from their final regression equations after stepwise analysis, in which the effect sizes are predicted from a number of study characteristics, including the amount of systematic language instruction in the treatment, the amount of systematic language instruction in the control group, and whether tutors were used to deliver instruction versus small groups or whole classrooms. These two sets of results differed in the weights that were applied in the analysis. The results from their Table 6 used weights (*WGT*) that gave equal representation to each study, where *k* represents the number of records contributed to the database by each study: $WGT1 = 1 / k$.

The results from their Table 7 used compromise weights (*WGT*3), or weights that resulted from multiplying equal representation weights with optimal weights (Hedges & Olkin, 1985, p. 110), which are a function of the sample size and effect size from each study. See Equations 1 and 2 and Camilli et al.'s (2003) Table 4 for a description of these weights:

$$WGT3 = WGT1 \times WGT2 \quad (1)$$

where

$$WGT2 = \left(\frac{n_{TOT}}{n_T n_C} + \frac{d^2}{2(n_{TOT} - 2)} \right)^{-1} \quad (2)$$

For both analyses, the regression weight associated with *TP2* is the parameter estimate of interest. This parameter represents the difference between (a) the effects that compare systematic phonics with a no phonics control and (b) those that compare some phonics with a no phonics control, which, in the context of Figure 1, is an estimate of *c*. However, in these models, the effect size estimate has been adjusted for the presence of other study characteristics; the systematic phonics effect has been corrected for the presence of systematic language activities in the same treatment or for treatment delivery through tutoring. The effect is to make the average of the *a* effects and the *b* effects closer to each other. When equal representation weighting is used (i.e., *WGT*1), the best estimate of *c* is .241; when compromise weighting is used (i.e., *WGT*3), the best estimate of *c* is .188. These estimates are lower than the NRP estimate but should not be directly compared to it because the NRP estimate is an average of the *a* and *c* effects.

In the same vein, Camilli et al. (2006) presented a third estimate of *c* in the context of a multilevel model that permitted control for the dependencies among effects from the same study. The parameter estimates from this model indicated that the effect of moving from some phonics to systematic phonics was $d = 0.123$ when other study characteristics and the dependency of effects within studies were controlled. Again, this is an estimate of the *c* effect adjusted for study moderators and cannot (and should not) be directly compared to the NRP unadjusted average of a mixed set of *a* and *c* effects of $d = 0.41$.

Another approach to estimating the effect size obtained by the NRP from the Camilli et al. (2003) recoded and reanalyzed data would be to use the parameter estimates from the regression model in which compromise weighting is used. Our best estimate of the *a* effects may be obtained by adding the constant from this model (.349) to half of the regression weight (because of the effect coding that was used) for *TP2*, or .188. We obtain an estimate for the effects of systematic phonics when compared to a no phonics control of .349 + .094, or $d = 0.443$. We can then use the parameter estimate for *TP2*, or .188, as our best estimate of the size of the *c* effect, or the difference between groups that received systematic phonics and those that received some phonics. If we assume that we had equal numbers of *a* effects and *c* effects to average, the best estimate of the overall average effect (without weighting by sample size) would be $d = 0.316$, or the average of .443 and .188. If there were more studies that evaluated the effect of systematic phonics compared to a no phonics control, the average would be closer to .443; if more studies evaluated the effect of systematic phonics versus a some phonics control, the average would be closer to $d = 0.188$. Although this average $d = 0.316$ is smaller than the $d = 0.41$ reported by the NRP, remember that this estimate represents an effect that has been adjusted for other study variables, such as the inclusion of language activities and tutors, so it should be smaller because the NRP estimates did not adjust for these moderators.

Hypothesis 2

Method

Database. To test the second hypothesis that combining different instructional activities was, on average, more effective, we used the Camilli et al. (2003) selection of studies and their coding of the NRP database and ran the regression model using the multilevel approach in Camilli et al. (2006). The analysis took into account the clustering of the data or the nonindependence of effects from the same study. It was not our intent to carry out a full multilevel model analysis with estimates of both fixed and random effects, but to replicate the regression analyses carried out by Camilli et al. (2006) on the modified database. Camilli et al. (2006) extended the work done in the 2003 reanalysis by recoding all of the studies for the level of systematic language activities in both the treatment and control groups using a 3-point rubric where a code of 0 indicated *no literacy activities*, 1 indicated *some literacy activities*, and 2 indicated *multiple language activities*.

Procedures. We deleted 25 effects from the total of 224 effects used in their analysis where there was neither some phonics nor systematic phonics instruction in the treatment group because we were interested in modeling treatment effects for phonics instruction and not for treatments that included no phonics. To facilitate interpretation of the regression parameter estimates, we recoded the *TP* variable into *TP_di*, where 0 represented *some phonics instruction* and 1 represented *systematic phonics instruction*. We agreed that controlling for the presence of other instructional characteristics in both the treatment and control groups might help explain some of the heterogeneity found in the NRP study and would also potentially yield a less biased average effect.

Other than these two modifications, we set up the multilevel model analysis in the same way as Camilli et al. (2006) using the SPSS code for their analysis, which was available from the Web site for Camilli et al. (2003). We translated their code into SAS Proc Mixed code and then reran their analyses on their original variables and the full set of effects to verify that we were running the same multilevel model. We then ran this model on the smaller, recoded data set (from which the 25 superfluous effects had been deleted) and calculated the predicted effect size values from the resulting model for all combinations of our predictor variables: the amount and systematicity of phonics instruction in the treatment group (*TP_di*), the amount and systematicity of language activities also present in the treatment group (*TL2*), the amount and systematicity of language activities in the control group (*CL2*), and whether or not the intervention was delivered one on one versus in small groups or classrooms (*Tutor*). The coding for *TL2*, *CL2*, and *Tutor* were included in the published database and were not changed.

Results

The estimates of the fixed effects from our multilevel regression analysis are presented in Table 1. All of the treatment characteristics included in the model significantly predicted unique variance in the effect sizes. To evaluate the practical impact of these various treatments options alone and in combination, we combined, as in any standard regression-based prediction model, the coded values for the predictors with the regression weights (see Table 1) to calculate predicted mean effects for each combination of predictor values. We then sorted the predicted values of the effect sizes from smallest to largest and present the mean predicted effect sizes along with the study characteristic codes in Table 2. Note that these analyses are based on the literature search by the NRP and the recoding of data by Camilli et al. (2006) and are therefore restricted to this body of studies.

Table 1
Fixed Effects From the Multilevel Model Modified From Camilli et al. (2006)

Effect	Estimate	*SE*	*t*	*p*
Intercept	.306	.106	2.88	.007
Tp_di[a]	.183	.089	2.06	.042
TL2[b]	.211	.106	1.98	.049
CL2[c]	−.213	.079	−2.71	.008
Tutor[d]	.424	.158	2.69	.008

Note. $df = 36$ for intercept; $df = 158$ for each of the other effects.
[a] Phonics in the treatment group: 0 = *unsystematic,* 1 = *systematic.* [b] Language activities in the treatment group: 0 = *none,* 1 = *some,* 2 = *systematic.* [c] Language activities in the control group: 0 = *none,* 1 = *some,* 2 = *systematic.* [d] Intensity of instruction: 0 = *classroom- and small-group level,* 1 = *one on one.*

Each row of Table 2 represents a potential combination of treatments and the expected effect size for that combination of treatments based on existing data. For example, the lowest *d* (−0.121), which may be found in the first row of the table, is the predicted value of *d* when there is some phonics in the treatment group, no language activities in the treatment group, no tutor, and systematic language activities in the control group. Under this condition a negative effect size favoring the control group is not surprising given the unique positive effect of systematic language activities that, in this comparison, are present in the control group but not the treatment group. Next, note the line in Table 2 in which all of the predictor values are 0 and the predicted effect size is $d = 0.306$. This line represents some phonics in the treatment group, no language activities in the control or treatment groups, and no tutoring. The value of predicted *d* for this scenario is equal to the intercept from our model. If we then move to the line where the *TP_di* variable is 1 and all other predictors are 0, the predicted value of *d* is 0.488. Thus, if we compare a treatment group that receives systematic phonics, no tutors, and no additional language activities to a control group that also has no phonics or systematic language activities and is taught in small groups or whole classrooms, we can hypothesize that the treatment group would perform .488 standard deviations higher than the control group on the outcome assessment. This effect is simply the intercept plus the effect of moving one unit on the predictor while holding the other predictors constant.

Next, examine the row in Table 2 in which there is systematic phonics instruction plus 1:1 tutoring. The predicted effect is $d = 0.913$, a very large effect. If a practitioner already had a systematic phonics program in place, this table could be used to get an idea of the potential effect of adding tutoring or additional literacy components. In the first case, the hypothesized improvement would be approximately .425, or the difference between .913 and .488, not coincidentally the value of the beta weight for tutoring in this analysis. The hypothesized effect of adding both tutoring and systematic language instruction to an existing systematic phonics

Table 2
Predicted Values Based on the Parameter Estimates From Table 1 and Coded Values for the Predictors

Predicted value	*TP_di*[a]	*TL2*[b]	*CL2*[c]	*Tutor*[d]	No. of effects	*k* (No. of studies)
−0.121	**0**	**0**	**2**	**0**	**12**	**3**
0.062	**1**	**0**	**2**	**0**	**12**	**4**
0.090	**0**	**1**	**2**	**0**	**7**	**1**
0.092	0	0	1	0		
0.273	1	1	2	0		
0.275	1	0	1	0		
0.301	0	2	2	0		
0.303	0	1	1	0		
0.305	0	0	2	1		
0.306	**0**	**0**	**0**	**0**	**30**	**6**
0.484	1	2	2	0		
0.486	1	1	1	0		
0.487	1	0	2	1		
0.488	**1**	**0**	**0**	**0**	**85**	**17**
0.514	0	2	1	0		
0.516	0	1	2	1		
0.517	0	1	0	0		
0.518	0	0	1	1		
0.697	1	2	1	0		
0.698	1	1	2	1		
0.700	**1**	**0**	**1**	**1**	**11**	**2**
0.727	0	2	2	1		
0.728	0	2	0	0		
0.729	0	1	1	1		
0.731	**0**	**0**	**0**	**1**	**7**	**2**
0.909	**1**	**2**	**2**	**1**	**3**	**1**
0.910	**1**	**2**	**0**	**0**	**3**	**1**
0.911	**1**	**1**	**1**	**1**	**4**	**1**
0.913	**1**	**0**	**0**	**1**	**17**	**5**
0.940	0	2	1	1		
0.942	0	1	0	1		
1.122	1	2	1	1		
1.124	1	1	0	1		
1.153	**0**	**2**	**0**	**1**	**4**	**1**
1.335	**1**	**2**	**0**	**1**	**4**	**1**

Note. Boldface indicates actual data from the meta-analysis. Other rows are extrapolations based on the regression equation.
[a] Phonics in the treatment group: 0 = *some,* 1 = *systematic.* [b] Language activities in the treatment group: 0 = *none,* 1 = *some,* 2 = *systematic.* [c] Language activities in the control group: 0 = *none,* 1 = *some,* 2 = *systematic.* [d] Intensity of instruction: 0 = *classroom- and small-group level,* 1 = *one on one.*

program based on the body of studies in the NRP report as recoded by Camilli et al. (2003, 2006) is $d = 1.34$.

Note that the nonbolded rows in Table 2 represent combinations of the predictor variables for which there are no studies in the extant literature from which to estimate effects. The predicted *d* in this case is an interpolation based on the data in hand. We have bolded the rows that represent predicted effects based on actual effects in the meta-analysis and have included the number of effects and the number of studies for each of these rows. Obviously, we would have more confidence in estimating effects based on more substantial data than on only a few studies or on interpolation.

The estimates from Table 2 can be used to generate hypotheses of anticipated results that need testing in an experimental study. Caution should be exercised when comparing the effects of systematic phonics instruction to those for tutoring or language activities from Table 2. The set of studies identified by the NRP was the result of a search methodology designed to find all possible studies that evaluated systematic phonics instruction using a randomized design or a high-quality quasi-experiment. This search was not designed to locate and include all possible studies evaluating the effect of using tutors to deliver instruction or of using language activities. Confidence in the magnitude of the systematic phonics effect is stronger because the NRP sought and identified the population of published studies.[1] The sample of tutoring and language activity effects contained within these studies is probably not the population of all such published effects and is also not a random sample from the population of all such studies; as a result, it is subject to selection biases. The effect on our results is that we

[1] Meta-analysis of the effect of group size do not support the greater efficacy of small-group tutoring in sizes of 1:1 versus 3:1 (Elbaum, Vaughn, Hughes, & Moody, 2000). Camilli et al. (2003, 2006) and the NRP compared 1:1 tutoring with the combined effects of small-group and classroom instruction.

should have less confidence in the precise magnitude of these effects or our ability to estimate the additive models laid out in Table 2. These cautions also apply to the conclusions reached by Camilli et al. (2003, 2006) and Hammill and Swanson (2006) to the extent that they are based on study characteristics other than the effect of systematic phonics.

Hypothesis 3

Method

Procedures. Using the effect size estimates from the NRP report and different estimates of the incidence of reading difficulties, we used a variant of a binomial effect size display (Rosenthal & Rubin, 1982) to evaluate the practical significance of effect sizes of different magnitudes under different base rates of struggling readers. In this variant we used realistic base rates rather than the 50% base rate in the original formulation of binomial effect size displays.

Statistical procedures. To create these displays, we took advantage of the fact that the relation between two dichotomized variables may be summarized with the phi coefficient, which is equivalent to the correlation coefficient effect sizes used by Hammill and Swanson (2006). We created a 2×2 table to represent the relation between a phenomenon of interest with two levels (high school dropout vs. high school completion) and a treatment with two levels (treatment vs. control). We placed the existing base rate for the phenomenon into the control level of the treatment and then, for a given level of phi, calculated the frequency of the phenomenon in the treatment group required to obtain that level of phi. Table 3 contains a table of symbolic frequencies to be used in conjunction with the formula for phi to solve for the unknown frequency for a given base rate and given phi:

$$\Phi = (ad - bc)/\sqrt{\text{efgh}}.$$

Using a through h in this equation, if we assume a value for c based on our knowledge of the base rate of a given phenomenon, and set g and h equal to each other, we can solve for the value of d that would correspond to any given phi value. For our examples, we chose total frequencies within the treatment and control to be equal ($n = 10{,}000$ each) and large enough to allow us to calculate frequencies in the treatment group in whole numbers. We used our assumed base rate to obtain the frequencies a and c. If the base rate of high school noncompletion were 10%, we would set a equal to 9,000 and c equal to 1,000 and solve for the value of d associated with a given level of phi.

Hypothetical scenarios. We created three hypothetical examples addressing the potential impact of even small effects within the context of reading instruction. In Scenario 1, we used the rate of at-risk readers in early intervention programs, commonly (but somewhat arbitrarily) estimated at 20% (Torgesen, 2000), as the characteristic we would like to change. In Scenario 2, we reduced the effect size in Scenario 1 and lowered the base rate to the high school noncompletion rate of 10% (Laird, DeBell, & Chapman, 2006) as the phenomenon we would like to reduce. In Scenario 3, we maintained the small effect size in Scenario 2 but increased the base rate to represent the number of struggling readers in the United States, which was cited as up to 40% by Hammill and Swanson (2006; see Shaywitz, 2004).

Table 3
Example of Symbolic Frequencies to Be Used With the Formula for Phi to Solve for the Unknown Frequency for a Given Base Rate and Phi

Variable	Control	Treatment	Total
High school completion	a	b	e
Noncompletion	c	d	f
Total	g	h	$a + b + c + d$

Results

In the first scenario, consider $d = 0.48$, $r = .23$, an effect size called "small" by Hammill and Swanson (2006). This effect was chosen because it was the effect for typical achievers in first grade from the NRP report. In Scenario 1, assuming a sample of 20,000 (half treatment, half control) and a base rate of 20% of at-risk readers who participate in a reading program with a predicted $d = 0.48$, $r = .23$, Table 4 shows a cross-tabulation of treatment and at-risk classifications. An intervention that results in a phi coefficient of .23 between the treatment and classification is consistent with an expected frequency of at-risk readers in the intervention group of 5%, a substantial decrease from 20%.

Similarly, consider an intervention that yields an effect size $d = 0.23$, $r = .11$, with respect to the high school dropout rate (see Table 4). We selected this effect size, the smallest significant effect reported in the NRP meta-analysis, because not only was it labeled small by Hammill and Swanson (2006), but it is also close to the Cohen (1988) heuristic for a small effect size. Assuming a 10% high school dropout rate without an intervention (Laird et al., 2006), an intervention that correlates with dropout status at the $r = .11$, $R^2 = .012$, level could result in a reduction in the dropout rate from 10% to 4.5%.

Finally, consider an intervention for which we expect a small effect of $d = 0.23$, $r = .11$, and set our base rate for struggling readers at 40%. Adjusting the base rate so that it is higher should reduce the potency of an intervention with a small effect size. Even with an intervention with this small effect size, we could hypothesize a reduction in incidence of poor readers from 40% to 30% (see Table 4).

General Discussion

There are three major conclusions from our assessment of Camilli et al. (2003, 2006) and Hammill and Swanson (2006). First, in terms of the first hypothesis, the conclusions of the NRP report are not contradicted by the two reanalyses of Camilli et al. (2003, 2006). The comparisons by Camilli et al. (2003, 2006) ask questions that are different from the primary question asked by the NRP, but the results of the two sets of analyses can be reconstructed to yield comparable effect sizes for the effects of systematic phonics versus either unsystematic phonics or no phonics controls when the same study parameters are estimated. When the effect sizes have been adjusted for study moderators, as in Camilli et al. (2003, 2006), the estimates are expected to be lower than the NRP estimate because the other treatment characteristics tend to improve performance (i.e., are positively related to the effect size)

Table 4
Binomial Effect Size Display for 10,000 Students in Scenarios 1, 2, and 3 After Intervention

Group	Classification	No intervention	With intervention
	Scenario 1		
At-risk readers[a]	No	8,000	9,500
	Yes	2,000	500
	Scenario 2		
High school dropouts[b]	No	9,000	9,550
	Yes	1,000	450
	Scenario 3		
Struggling readers[c]	No	6,000	7,000
	Yes	4,000	3,000

Note. Scenario 1: base rate = .20, $d = 0.48$, $r = .23$; Scenario 2: base rate = .10, $d = 0.23$, $r = .11$; Scenario 3: base rate = .40, $d = 0.23$, $r = .11$.
[a] Incidence of at-risk readers reduced from 20% to 5%. [b] Incidence of high school dropouts reduced from 10% to 4.5%. [c] Incidence of struggling readers reduced from 40% to 30%.

and because the vectors of study characteristics are not orthogonal to one another. If the estimates are directly compared, the key is to estimate the same effect size across the different studies. An effect size is always a comparison of different conditions with one another (see Figure 1). The NRP and Camilli et al. (2003, 2006) estimated effects from different comparisons.

Second, Camilli et al. (2003, 2006) showed that both phonics and language instruction, as well as tutoring, impacted reading outcomes. They questioned whether it was accurate to assume that phonics instruction was the core of many of the programs represented in the NRP database, especially interventions in the control groups. These findings support the NRP contention that reading programs need to be comprehensive by providing estimates of the unique contributions of different factors that moderate the impact of phonics. To get larger effect sizes, reading instruction must be comprehensive and contain literacy components beyond phonics, especially when the diversity of students in classrooms and local instructional contexts are considered. Thus, in terms of the second hypothesis, reanalyses of the NRP report show that larger effects may occur in association with combining different components of reading instruction. Within the modified NRP corpus of studies, systematic phonics was associated with larger effects than no or some phonics, but the effects tended to be small to moderate and variable across studies, and much of the explainable variability in effects (i.e., that not associated with sampling error) was not accounted for by the intervention characteristics included in the model. The addition of literacy components that presumably support fluency and comprehension, as well as tutoring, was associated with larger effects. However, larger effects were associated with systematic phonics, regardless of the levels of systematic language activities and tutoring. Within the modified NRP corpus of studies, the largest effects were associated with the combination of systematic phonics with additional language and literacy activities and one-on-one tutoring.

Third, whether we characterize $d = 0.41$ as moderate or small, the evaluation of Hypothesis 3 shows that even small effect sizes may be of sufficient magnitude that they could be associated with significant reductions in the incidence of reading problems. Hammill and Swanson (2006) utilized Cohen's rules of thumb for planning studies as thresholds for determining the practical significance of effect sizes but did not take into account context and base rates in their dismissal of small effect sizes, essentially committing a Type II error. Although they were correct in suggesting that additional variability is unexplained, it is an inappropriate extrapolation to suggest that the amount of explained variability is not practically significant. In Scenario 1, with a medium effect and a medium small base rate, the hypothetical intervention reduced the incidence of struggling readers from .20 to .05, for a reduction of 75%. In Scenario 2, with a small effect and a lower base rate than Scenario 1, the hypothetical intervention reduced the incidence from .10 to .045, for a reduction in struggling readers of 55%. In Scenario 3, with the same small effect as Scenario 2 and a high base rate, the given effect size was associated with a reduction in struggling readers from .40 to .30, or 25%. These examples show that the impact of even small effect sizes may be practically important, especially when coupled with low base rates of the phenomenon of interest. The next step in placing these effects into a context involves computing the costs associated with delivering a given intervention with the benefits expected from moving some number of individuals from one category to another and comparing these costs with expected benefits. This step is outside of the bounds of this article, and, in fact, because costs and benefits are context dependent, the practical significance of a given effect size might be decided on a location-by-location basis. However, because reading instruction is routinely provided to students in schools, the costs in changing instructional emphases should be relatively small compared to the overall costs already in place for teaching children to read. The effect size and the base rates used in our examples are comparable to situations that exist with different school settings and interventions.

Altogether, these results support approaches to reading instruction that are more comprehensive and, for alphabetics, approaches that are more explicit and in which the knowledge is directly shared relative to those in which knowledge must be inferred by students. Whether these principles extend beyond just the effects of phonics instruction cannot be established from the NRP report or Camilli et al. (2003, 2006), although other reviews support more explicitness for fluency and comprehension (Pressley, 2005). In reaching this conclusion, we note that these pedagogical principles exist on a continuum and should not be dichotomized. In examining this continuum for instruction involving the alphabetic principle, it may be that the more important component is explicitness and the deliberate attempt to instruct the child as opposed to a scripted approach to phonics, especially if the child is at risk for reading difficulties or is struggling to learn to read. Indeed, both the NRP and Camilli et al. (2006) concur in estimating larger effects of systematic phonics for students who are struggling readers, findings supported by recent experimental studies that formally manipulate explicit instruction in relation to child characteristics. For example, Connor et al. (2007) found that more time in phonics instruction is beneficial to students weak in alphabetic knowledge; conversely, more time on comprehension instruction

leads to better outcomes in students weak in vocabulary instruction.

To illustrate the difference in scripted versus explicit, Torgesen et al. (2001) compared the efficacy of a highly scripted reading program with a clearly defined scope and sequence with an approach that taught the alphabetic principle explicitly but spent more time reading and writing in context. There were no significant differences in outcomes for a group of elementary school children with severe reading disabilities (see also Wise, Ring, & Olson, 2000). Mathes et al. (2005) compared two comprehensive small-group tutorial interventions based on (a) a direct instruction model with a scripted lesson plan and well-developed scope and sequence with use of decodable text; and (b) a guided reading intervention in which instruction in the alphabetic principle was explicit (i.e., based on a plan for introducing phonics elements and in which the information was directly presented to the child) and done for about 20% of the instructional period, but unscripted and with the use of leveled texts instead of decodable texts. No major differences in reading outcomes for first graders at risk for reading difficulties were apparent when these two comprehensive programs were compared.

These examples show that the explicitness of instruction may be more important than systematic, scripted lessons in accounting for the effect of systematic phonics. Creating a scope and sequence, using decodable text, and engaging in other ways of systematizing instruction make instruction explicit, but explicitness can be achieved in other ways. Where a teacher operates on the instructional continuum may depend on factors like preparation, experience, the base rate of struggling readers, the school context, and related factors. However, teachers need to be intentionally clear about how the alphabet relates conventionally to sound segments in speech. The supporting materials that are used may vary depending on teacher and student knowledge and skills.

In contrast to the seemingly endless political and ideological commentaries about the purposes and findings of the NRP study, Camilli et al. (2003, 2006) and Hammill and Swanson (2006) have advanced the field by focusing on the data and the need for replication and confirmation of the NRP findings. Adjudication of the issues raised by the NRP report through attempts at replication and continued experimentation can help move the field beyond the simplistic instructional dichotomies that have plagued theory and instruction on reading toward richer and more complex approaches that will enhance reading proficiency for all children. Nonetheless, our analyses of Camilli et al. (2003, 2006) and Hammill and Swanson do not support the belief that the NRP misrepresented their findings or misled policymakers and the educational community (Allington, 2006). The NRP relied on empirical synthesis (meta-analytic methods) for the interpretation of a large body of research. The NRP report explicitly stated the criteria for including studies in the meta-analysis and was subjected to peer review prior to its release to the public.[2] Our reanalysis of the NRP findings confirm their conclusions concerning phonics instruction but must be understood in the context of the need for comprehensive approaches to reading instruction. As the NRP (NICHD, 2000, p. 2-97) stated, "Phonics instruction is never a total reading program," and it "should be integrated with other reading instruction."

These conclusions lead to what we believe should be the reading community's vision of an effective reading program. That is, comprehensive instruction involves explicit instruction in the alphabetic principle, explicit instruction in comprehension and vocabulary, and active engagement of the child to develop fluency (Pressley, 2005; Snow et al., 1998). Few in the reading community would disagree that this task is arduous and hardly begun, and the next step is advancing beyond the findings of the NRP and other consensus reports. 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 effect sizes may be larger than for any of these components in isolation. That is the important message of the NRP report, Camilli et al. (2003, 2006), and Hammill and Swanson (2006). Although it seems difficult to move beyond the historic dichotomy of reading instructional approaches, it is time to embrace comprehensive approaches to reading instruction and work toward determining how to integrate different components of reading instruction into classroom practice so that the diversity of students and their individual needs can be addressed.

[2] Camilli et al. (2006, p. 30) were in error when they indicated that the NRP report was not subjected to peer review prior to its release (P. McCardle, personal communication, February 21, 2007). In suggesting that the NICHD change procedures for producing meta-analyses, there is a misunderstanding. Consensus reports at the National Institutes of Health are usually done by the Office of Medical Applications of Research (OMAR), which convenes panels of scientists to produce consensus reports. Although the NRP was congressionally mandated, it preceded OMAR in deciding to use meta-analysis. Procedures for conducting syntheses, including the use of meta-analysis, are determined by the specific committee, not the National Institutes of Health or OMAR.

References

Allington, R. L. (2002). *Big brother and the national reading curriculum: How ideology trumped evidence.* Portsmouth, NH: Heinemann.

Allington, R. L. (2006). Reading lessons and federal policy making: An overview and introduction to the special issue. *Elementary School Journal, 107,* 3–15.

American Federation of Teachers. (1999, June). *Teaching reading is rocket science: What expert teachers of reading should know and be able to do.* Washington, DC: Author.

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., & Smith, M. L. (2006). Meta-analysis and reading policy: Perspectives on teaching children to read. *Elementary School Journal, 107,* 27–36.

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

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science, 315,* 464–465.

Cooper, H. (2005). Reading between the lines: Observations on the report of the National Reading Panel and its critics. *Phi Delta Kappan, 86,* 456–461.

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

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps chil-

dren learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36,* 250–287.

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92,* 605–619.

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,* 37–55.

Foorman, B. R., Schatschneider, C., Eakin, M. N., Fletcher, J. M., Moats, L. C., & Francis, D. J. (2006). The impact of instructional practices in Grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology, 31,* 1–29.

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.

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

Laird, J., DeBell, M., & Chapman, C. (2006). *Dropout rates in the United States: 2004* (NCES 2007–024). Retrieved March 20, 2007, from http://nces.ed.gov/pubsearch

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

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). An evaluation of two reading interventions derived from diverse models. *Reading Research Quarterly, 40,* 148–183.

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.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). Fort Worth, TX: Holt, Rinehart, & Winston.

Prentice, D., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164.

Pressley, M. (2005). *Reading instruction that works: The case for balanced instruction* (3rd ed.). New York: Guilford Press.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Santa Monica, CA: RAND.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2,* 31–74.

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

Shanahan, T. (2004). Critiques of the National Reading Panel report: Their implications for research, policy, and practice. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 235–265). Baltimore: Brookes.

Shaywitz, S. E. (2004). *Overcoming dyslexia*. New York: Knopf.

Snow, C. E., Burns, S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford Press.

Steering Committee of the Physicians' Health Study Research Group. (1989). Final report on the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 321,* 129–135.

Swanson, H. L., Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcome.* New York: Guilford Press.

Thorndike, R. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Torgesen, J. K. (2000). Individual responses in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice, 15,* 55–64.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33–58.

Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the *Journal of Counseling & Development. Journal of Counseling & Development, 82,* 107–110.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51,* 473–481.

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

Wise, B., Ring, J., & Olson, R. K. (2000). Individual differences in gains from computer-assisted remedial reading with more emphasis on phonological analysis or accurate reading in context. *Journal of Experimental Child Psychology, 77,* 197–235.

Received April 16, 2007
Revision received August 11, 2007
Accepted August 30, 2007 ■