Some Comments of a Report by C. Torgerson, G. Brooks, and J. Hall titled "A Systematic Review of the Research Literature on the Use of Phonics in the Teaching of Reading and Spelling.

Diane McGuinness

Recently, basic-skills website featured a summary of a review of the reading research by Torgerson et al (Research Report no. 711). The 85 page document is linked to this summary. It is, in effect, purported to be a reworking of the National Reading Panel's report in the US (2000), or more accurately, the reading committee chair's report of the same data (Ehri et al, 2001).

This report was commissioned by the DfES and funded by them. The committee, headed by Professor Greg Brooks, was "supported and advised" at "each stage of the review with helpful comments and suggestions." This is a dense document, with numerous tables and appendices, arcane discussions of statistical minutiae and issues regarding experimental design, etc., all to the end (it appears) of drawing a vague set of conclusions which lead the reader to believe that synthetic phonics programmes have not been proven to be effective beyond other methods by any margin sufficient to be trustworthy. The reality is, that every statement under the heading "key findings" is incorrect or seriously compromised by the true facts.

Before we look in detail at a few of this group's claims, we need to look first at the original study itself. In 2000, the National Institute of Child Health and Development, under the auspices of the US National Institutes of Health, published a remarkable document which became known as the National Reading Panel report or NRP for short.

In this report every category of reading instruction (apart from spelling per se) was investigated by a series of database searches. For reading instruction, the committee unearthed 1,072 research reports. Using a set of basic criteria to ensure scientific validity, the committee reduced this number to 75, and by more careful screening to 38 which included a total of 66 comparisons of reading methods.

The focus from the outset was on instruction identified as "synthetic phonics" which was described as "sounding out and blending of words versus programmes teaching students to decode using large subunits of words, versus miscellaneous programmes." A subsequent definition included the phrase "sounding out of phonemes" (as opposed to other units of sound. As a general rule, the control programmes were some type of whole word learning, either with lists of sight words, or with guessing via context cues, etc. or a combination of the two. In rare cases, synthetic phonics was compared to analytic phonics.

As the report proceeded, terminology shifted depending upon who was doing the writing. "Synthetic phonics" turned into "systematic phonics" with more or less the same definition. Overall, there seemed to be a good deal of confusion between writers as to whether "synthetic phonics" meant learning the sounds of the individual letters, letter

digraphs, and phonograms –(as many as "90 sounds" were advocated in one study) *or* whether it meant learning the 40+ phonemes of the English language, the "basic code spellings" for each of them, and spelling alternatives for these sounds.

The term "synthetic phonics" lacks what, in scientific terminology, is known as an "operational definition." When something is operationally defined, this means it has parameters which can be measured with sufficient mathematical rigour that these measures can be relied upon to define the concept. This makes it possible to replicate a study with greater accuracy.

Unfortunately, we are not there yet, and the NRP Reading Committee used the following criteria to screen studies into their database for use in a meta-analysis (a statistical tool for combining results by converting the data to standard scores).

> The study must use an experimental or quasi-experimental design with a control group.
> Studies must appear in a refereed journal (peer review) after 1970.
> Research must address the central question as to whether "systematic phonics" instruction improves reading performance more than other types of instruction.
> Reading has to be measured as an outcome.
> Statistical analysis must be sufficient to calculate effect sizes (which means standard deviations are included, or can be estimated).

As anyone can see, these screening criteria do not solve the operational definitions problem. And quite apart from this problem, there are enormous difficulties with meta-analysis research. Without great care researchers can combine apples and oranges, in which the variables are too dissimilar to be combined. This is an issue of logic, and requires great care. For example, the NRP combined the data across ages, across types of instruction (single child, small groups, all children in a classroom), across the number of hours of instruction (days, weeks, months), quite apart from the range of programme types they believed fit the designation: "synthetic" or "systematic." These included Orton-Gillingham with its emphasis on letter names as well as sounds and a strong focus on memorizing "syllable types," onset-rime/analogy programmes, and more bread-and-butter synthetic phonics programmes that were either well thought out or not.

The Torgerson et al analysis makes many of these same mistakes, and others as well.

There is no doubt that the single most important distinction between data sets that don't combine is the difference between remedial tutoring and early classroom instruction. Remedial tutoring is generally one-on-one (and research ***must*** include a control group learning by some other method), and remedial methods are vastly different to early classroom methods. In addition, remediation is used almost exclusively for older children who have failed in the classroom, and so these children are much older than beginning readers. In addition, they have bad habits that the beginning readers don't share. For all these reasons (age, method, habits, matched control groups), studies on remedial tutoring and early classroom instruction cannot be compared directly, and

certainly should not be included in the same meta-analysis. This precaution was not taken by the NRP committee nor by Torgerson et al.

This fact alone, invalidates the general conclusions from both studies. Fortunately, the NRP provided a number of tables with extensive descriptive and numeric content. This makes it possible to look at each individual study in some detail, and to compare and combine the data in different ways.

By the simple act of separating the studies into these two categories, and computing effect sizes for each one separately, one can see the huge differences in outcome. Remedial reading programmes (as a group) fail dismally with effect sizes of .27. The weakest among them are studies using Orton-Gillingham methods (10 cases, E.S. = .23). When the remedial studies were removed from the data pool, and the classroom studies compared, the effect size rose dramatically to .55, an improvement of over ½ a standard deviation.

And there is more. When the results for onset-rime/analogy based "phonics" programmes were compared to those for phoneme-based synthetic phonics instruction, there was another parting of the ways. Rime/analogy programmes were singularly unsuccessful, producing an effect size of .28. The pooled data for the phoneme-based programmes on their own produced an effect-size shot of around 1.0 (one-standard deviation advantage over the comparison groups). This effect is very large, and it is reliable.

In my recent book, *Early Reading Instruction* (2004), I attempted to get beyond the global and semantically confusing labels by supplying some missing operational definitions derived from the basic reading research, as opposed to the applied reading research. The key to this analysis was a series of "time on task" studies in which individual children were observed for many hours over a period of months. There are several studies of this type in the literature and every one of them reported identical results.

Success on standardized reading tests (both decoding accuracy and reading comprehension) at the end of the school year are directly correlated to the time spent on the following tasks:

- learning the phoneme-symbol relationships of our alphabet code
- time spent learning to sound out words phonemically and blending these sounds into words
- time spent writing letters, words, sentences, and stories
- time spent reading text which can be sounded out and blended

Other tasks, by contrast, had a strong negative impact on learning to read, so much so that the amount of time spent on these tasks produced declining reading test scores compared to test norms. The correlations between time on task and reading test scores were minus .80 or even higher. These are the tasks that require either 1) memorizing words by sight or guessing words in context, or 2) time spent on vocabulary lessons and listening to the

teacher read. The first activity is strongly detrimental to learning to read, and the second two activities are simply neutral (have no effect other than to delay learning how to master the code).

In addition to the time on task studies, other work showed that children made far greater progress in learning to read if phonemes (and no other sound units) were the basis for learning the code. Children did much better if they learned phoneme/letter correspondences from the outset rather than learning phonemes in isolation.

Based on a number of outstanding research studies on these issues carried out over a period of 45 years, plus new knowledge about how writing systems developed and how they work, it was possible to set out a *prototype* for reading instruction. The prototype is, in effect, a set of "operational definitions" for particular tasks that are predicted to produce a successful outcome for classroom reading instruction (and for remedial instruction as well). Fortunately, there are a number of classroom programmes that fit the prototype, though very few are found in the remedial sector, one reason why it is lagging so far behind.

When I applied the criteria of the prototype to the studies reviewed by the NRP, the programmes with the best fit were far and away the most successful, producing gains of 2 to 4 years above national norms, or above control groups.

These programmes are well known to English researchers and educators. There is no mystery here, no need to make statements like: "we have no research" to prove any method is better than another, or that "one cannot draw firm conclusions" based on the data so far, etc. This is simply untrue.

Successful programmes teach the 40+ sounds of the English language, their most common or likely spelling, plus a few or all alternatives spellings as well. In the UK, they are generally known as "synthetic phonics" programmes, whereas in the US, this is not always the case. Due to the confusion that has arisen over the term "synthetic", I proposed in my book, that we rename the programmes that fit the *prototype* -- "linguistic phonics." So far, this idea has not caught on.

The list of "linguistic phonics" programmes grows longer as time goes by. Early examples were Joyce Morris' *Phonics 44*, and the initial teaching alphabet (*i.t.a.*). In the US, there were the *Lippincott* and the *Lindamood programmes*, starting in the 1950s and 1960s. More recently there is Sue Lloyd's "*Jolly Phonics*," Johnston and Watson "*Fast Phonics First*," Ruth Miskin's "*Best Practice Phonics*," and C. and G. McGuinness' "*Phono-Graphix.*" Plus, there are a number of similar programmes with different names, and, a new family of programmes that were built on the prototype and designed for the home, school, and clinic (*Sound Reading Systems*). The latter programmes stress writing as a way of learning, and weave the entire spelling code into the lessons from the outset. This has been found to dramatically speed up learning.

Linguistic phonics programmes produce phenomenal (and consistent) results in the classroom as well as in the clinic. Pooling the data from five studies which used Jolly Phonics, or a similar programme (Fast Phonics First), produced combined effect sizes of .84 for reading, .89 for nonword decoding, 1.22 for spelling, and .73 for reading comprehension. Not only this, but these programmes had a huge impact on phoneme segmentation scores (E.S. = 1.65) compared to a control group. Moreover, effect sizes are much larger than those for programmes where phoneme awareness is taught on its own without any connection to letters.

In addition, the outstanding work of Johnston and Watson in following up their children for seven years, shows that, if anything, these gains continue to increase over time compared to national norms.

This brings us to the recent re-review of the NRP data identified at the top of the page. What, exactly, is its purpose? Does it provide us with any greater insight concerning the best way to teach reading, spelling, and reading comprehension?

The authors began with the same set of 38 studies reported by the NRP (specifically Ehri's own report, 2001). They take issue with several aspects of this approach, in particular the "methodological limitations" of Ehri's work. As a result, they started the process anew. "A total of 6114 potentially relevant studies were identified" and after the first screening, "101 potentially relevant studies were identified." There is no description of what is or isn't "potentially relevant" in these two cases. In any event, a comparison to the NRP report, shows that every study except three, was already in the NRP database, and had met the NRP screening criteria.

Here are some of the Torgerson team's stated criteria.

1. Screening must include non-peer review studies.
As it is difficult to publish research that gets non-significant results (which is true), therefore, Torgerson argued that focusing solely on peer reviewed research is invalid.

Anyone who knows this voluminous literature well, especially if they have encountered the vast wasteland of unpublished articles on ERIC (a database for anything and everything), will recognize that peer-reviewed material is absolutely paramount to doing research of this type. Otherwise one would have to play editor oneself, writing to authors for missing data, missing information, etc. in every case. This suggestion is simply preposterous.

2. Random assignment of students to experimental and control groups.
This topic received a great deal of attention, far more than it warrants, given the basic reality of how schools work. One must assume, for example, that the head, or assistant head "randomly" assigns children to teachers as they progress through the school. It isn't practical for researchers to come in and randomly assign them again. Applied research,

unfortunately, isn't carried out in a perfect world. The authors used this argument plus others to throw out a number of studies, thereby reducing the data set to 14 studies.

3. The authors categorically state that these 14 studies include comparisons between synthetic phonics and analytic phonics. I could only find one. Here is a list of studies that made it to the final 14, with a description of the children and the category of programmes that were used for the experimental and control groups.

Berninger et al 2003. (new) 7 year olds who failed to learn to read. Pull-out programme for pairs of children. Comparison is to training in word recognition versus reading comprehension.

Brown and Felton (1990) (NRP data). Used the Lippincott programme (synthetic phonics) for 6-year olds, classroom, in small groups (8 per group). Comparison group used look-and-say (sight word).

Greaney et al (1997) (NRP data). 8-year-old poor readers. Taught rime/analogy method. Controls had look-and-say.

Haskell et al (1992). Normal 6-year-olds. Classroom. Phoneme/onset-rime methods. Control taught whole word.

Johnston and Watson (2004). Normal, age 5 beginning readers. Synthetic/linguistic phonics. See above.

Leach and Siddal (1990). Normal, age 6. Direct Instruction (Reading Mastery) versus look-and-say.

Lovett et al (1989; 1990). Poor readers, age 10-11 years. Syn. Phonics/or analogies. versus whole language.

Martinussen and Kirby. Kindergarten, regular classroom. "Successive phonics" (synthetic phonics) versus whole language.

O'Connor and Padeliadu (2000). Age 6, poor readers. Group. Blending versus whole word learning. Sample size (6 per group) invalidates this study.

Skailand (1971). Kindergarten, phoneme/grapheme versus look-and-say. (unpublished). Later, the reader is told, that the standard deviation is larger than the mean, so this study is invalid – probably why it didn't get published.

Torgesen (1999/2001). Longitudinal study starting at age 5 (at risk) and a second study on children diagnosed LD, 8-10 years. Lindamood programme (pull out) versus classroom, unspecified.

Umbach (1989). Age six poor readers. Reading mastery versus look-and-say.

Here is the same problem of clinical populations working one-on-one or in small groups in a pull-out programme lumped together with children learning to read in the classroom. These two populations cannot be mixed to get any meaningful information.

The authors complained that Ehri et al failed to include research comparing synthetic versus analytic phonics, and had excluded 5 studies with this focus. The reader is told that "The current review includes an analysis of this topic." We never learn the outcome of this analysis, except to note that these 5 studies (judging from the list above) did not find their way into this paper either.

Later, they state that the studies screened into their database are either those which compare systematic phonics versus unsystematic/or no phonics, or synthetic phonics compared with analytic phonics. The only study reported that fits the latter category is the study by Johnston and Watson.

In conclusion, this paper shares the same major flaws as the NRP and a lot more besides. Mixing remedial and classroom programmes is invalid on a number of grounds. The reduced data set, which tossed out a number of very good studies indeed (i.e. Stuart's dockland study using Jolly Phonics is but one) is scarcely an improvement on the NRP report, or on its thoroughness and detail, which allows others to use the data more effectively.

When the data are assessed appropriately, the reading programmes that best fit the prototype, that fall most clearly within the ideal of the synthetic phonics programme that focus exclusively at the level of the phoneme and no other, larger units, is highly successful for beginning readers. The evidence is clear, robust, and comprehensive enough for this fact to be recognized.

The next stage of research needs to be focused on fine-tuning these outstanding programmes. We certainly do not need more mock research papers designed to "prove" their ineffectiveness.