Science Versus Basic Educational Research

by Siegfried Engelmann
February 2003

One goal of basic research in education is to identify the variables of effective instruction. As this pursuit has been conceived, however, its theoretical problems make it unlikely that the effort will provide a clear picture of instructional variables, their interactions, or the kind of teacher training that is implied by instructional variables.

Identifying Variables that Influence Student Performance

The most elementary fact is that some approaches to teaching a skill like beginning reading are more successful and replicable than others. What are the variables? How many of them are there? How do they interact?

There are at least four ways to go about finding answers to these questions, some of which are more efficient and scientifically sound than others.

> *Type 1 approach*. The most obvious way would be to ask and check. Because successful applications have been created by design, ask the designer of the successful program what the variables are that the design controls and why. The answers imply tidy, controlled experiments that involve systematic investigation of the designer's assertions.
>
> *Type 2 approach*. A related approach would involve fully implementing a successful instructional program and then systematically altering the details of it (one at a time, while trying to maintain the others as they were in the original program). The changes would be correlated with changes in student performance. If no difference results from a change, the dimension that was changed does not function as a variable (at least within the range of variation observed). A manipulation that results in improved student performance identifies a variable that was not well designed by the original program. A change that results in inferior student performance identifies a variable that was designed better by the original program than it was in the modified program. This approach would need clear descriptions of what constituted improved performance. Efficiency is an important variable. If the change resulted in improved performance but required three times the instruction of the original, the rubric for judging efficiency would have to compute the ratio of improved performance over the time to arrive at a reasonable overall judgment of the net "desirability" of the change.
>
> *Type 3 approach*. A less articulate technique would be to perform a "factor analysis," meta-analysis, or correlation of differences in student performance across various programs. The most serious problem with this approach is that it largely begs the question. The data will be sorted according to different factors. These are assumed to be the variables. If information about a particular factor is not provided in the information bank of the program, it cannot be correlated with student outcomes. So the factor analyses and meta-analyses are largely limited to the knowledge of the investigators, which is understandably not on the level of

> the designer or a technician. The analysis would certainly show whether one instructional program undoubtedly outperforms another; however, this information is revealed by a relatively simple analysis of the range of scores (given that the groups were comparable at the beginning of the experiment*).*
>
> *Type 4 approach.* An even less rigorous approach would be to identify features of successful programs using observations of groups being taught through a particular program. This approach is particularly crude because it relies on the observational prowess of the investigator. The investigator looks at two different programs, for instance, and notes how they are different. The investigator will more likely attend to those features that are highly visible, even though these variables may be relatively less important than some of the less visible ones that the investigator did not identify as variables. This approach has all the problems of Type 3 approach and, additionally, that of the quality of the data that are used to draw conclusions. A further problem, which this approach shares with Type 3, is that there are no rigorous procedures for interpreting the data. Because Type 4 approach uses a larger percentage of questionable data, the chances of distorted conclusions are potentially greater.

Type 1 approach is theoretically the most efficient and is adequately scientific. Its efficiency derives from the fact that it increases the probability of the investigator identifying subtle variables that are important. If something is subtle, such as the number of "sounds" in the initial words to be decoded, it may require hundreds of Type 2 or Type 3 approach manipulations to identify them. Yet the designer may be able to point out, for instance, that if all the initial words have three sounds, some children will later have trouble sounding out two-sound words because they will try to add a third sound to them (a rule based on what the poor program unintentionally taught). Also, if the initial words are in the order consonant-vowel-consonant (cat), children will have trouble identifying words that are in the vowel-consonant-consonant order (and) or the vowel-consonant order (it).

Type 2 approach is pristine and draws conclusions that are unquestionable. Its major problems are that it is laborious and it generates information that is restricted to the program that is analyzed. The logic of the design, however, unquestionably articulates the variables. If program X is highly successful and changes in details result in a performance difference, that performance difference is clearly a function of the changes.

The Type 2 approach format is superior to 3 and 4 because both variables that make a difference and the manner in which they are configured are identified unambiguously by making a simple hypothesis and testing it. If the investigator believes that initial work on phonemic awareness should present tasks that require the learner to decompose spoken words into phonemes, the hypothesis is testable, using the performance baseline of the successful program for comparison. When the data have been gathered, the investigator will have clear information that the hypothesis was resoundingly disconfirmed. The logic of a Type 2 approach is that a change in one or a small combination of program details either results in a performance change or it doesn't. This is a scientific basis for identifying variables that are not influenced by opinion.

Type 3 and 4 approaches may have a scientific basis; however, if this basis is used to identify those broad “policy and management” factors, like funding and class size, the approach seriously confuses correlation with causation.

Class size will influence performance; however, changing class size will logically have a differential effect on various programs. If the size of kindergarten classes were increased to the point that all programs would fail (say a size of 90 with no aides or helpers), all programs would tend to perform the same because all programs are now reduced to roughly the same performance level (poor). The greatest performance reduction occurs with the highly effective program; the one that suffers the least is the poor program. Therefore, the most sensible approach is to work only with programs that have substantial evidence of effectiveness and then systematically identify the effects by changing class size.

Another problem with Type 3 and 4 approaches is that it is often fashionable to draw conclusions that are illogical. For Type 3 approach, a legitimate conclusion based on data of common features of successful programs would be of the form, “If programs are successful, they have features A, B, C, D.” This order is implied because the independent variable was performance. The dependent variables were the features. The analysis distorts the relationship if it states (or implies), “If programs have features A, B, C, D (false independent variables), they will be successful (false dependent variable).” This order is not legitimate because the investigation did not search for programs that have features A, B, C, D, but rather programs that were successful.

The improper order suggests that the successful program will be created if the designer includes features A, B, C, D in the program. This suggestion has no logical or experimental basis. The conclusion has never been tested or rigorously demonstrated. In other words, it is a fabrication.

Note that these misuses are not necessary functions of Types 3 and 4 analyses. A Type 4 approach analysis may be used simply to judge whether program X is being implemented properly. If it is, it will have the features A, B, C, D. If some of these features are not present, it is not program X, but a poorly imitated facsimile.

In the same way there are conclusions that are legitimately generated by a Type 3 approach analysis. If all successful programs have features A, B, C, D, and if program Z does not have these features, it is fair to conclude that this program will not be as successful as the programs that have the features. However, the most direct reason (and best evidence) for concluding that program Z is not successful has nothing to do with features A, B, C, D, but rather with the student-performance data, which show it is not successful. This evidence is not usually available for programs that are “new,” which may be why the field of education tends to introduce a lot of new and untested approaches.

Experimental Recommendations

As noted above, by suggesting that a set of features will lead to an excellent program, Type 3 and 4 analyses have a serious logical problem, which is that although the *facts* the analysis uses are data-based, the *conclusions and recommendations* drawn from these facts are often speculative or “experimental” in nature. If the report’s recommendations imply that by assembling features A, B, C, D an excellent program would result, the recommendation would be an assertion about something that has never been experimentally verified in any form. The only possible way to determine the validity of the assertion would be to give various program developers instructions to create programs that had features A, B, C, D. The student performance would be evaluated and would determine the extent to which all program developers produced excellent programs. If all designers who followed the directions achieved excellent student performance, the recommendation would be valid. If only some of them produced excellent programs, the recommendation would not be valid and would require qualification.

Analytical Tendencies

Type 1 and 2 approaches have not been used to evaluate approaches for teaching skills like beginning reading, although these approaches are capable of yielding the most articulate information about the variables and their interplay. Combinations of Type 3 and 4 approaches have been used extensively to formulate “recommendations” of what should be done to teach content like beginning reading. The studies are typically guilty of all the logical problems and distortions noted above.

The first serious attempt to identify variables for beginning reading programs was *Becoming a Nation of Readers* (Anderson et al., 1984), written by the Commission on Reading, composed of what were considered to be the best researchers, and jointly conducted by the National Academy of Education, the National Institute of Education, and the Center for the Study of Reading at the University of Illinois. Although the analysis referred to some of the techniques and variables that characterized effective programs, the designers of these programs were not quoted or consulted on the variables, their interaction, or importance. No Type 1 or Type 2 approach investigations were cited (or conducted). None of the highly effective programs were identified or recommended. Instead, various research studies (ranging from those that were well designed to those that were scientifically wanting) and viewpoints that had no particular data base were liberally presented in the discussions. All recommendations were advertised as being consistent with the best available evidence and that they would improve reading instruction. In general, these assumptions were verified; however, in some cases, the variables that the Commission identified were not data-based and not features of highly effective approaches.

For instance, in discussing when reading should begin, the Commission makes this observation:

> There is a wealth of evidence that children can benefit from early reading and language instruction in preschool and kindergarten.[26] Available data suggest that the best short-term results are obtained from programs that can be characterized as formal, structured, and intensive. [27] (p. 29)

A reference note (27) that documents this final assertion refers to the Becker and Engelmann technical report*, Analysis of Achievement Data on Six Cohorts of Low-income Children from 20 school districts in the University of Oregon Direct Instruction Follow Through Model* (1978). This report covered over 25,000 students. The Direct Instruction model outperformed all other models in Follow Through in all subjects. It had 30 educationally significant advantages in reading over the Follow Through comparison groups. (An educationally significant outcome was one that produced at least 1/4 standard deviation difference between Direct Instruction and the comparison groups.) Based on this data it would seem that whatever practices were used by the Direct Instruction model would carry more weight than those of failed models.

The next observation the Commission makes about reading in kindergarten, however, is: ". . . though whether these programs have greater long-term benefits is less clear. Good results are also obtained with informal, though not haphazard, programs." One of the references for this assertion was a study by Schweinhart and Weikart (1980) that involved 54 children in three different preschool programs (18 in each), followed up ten years after the intervention. The Weikart model (High/Scope) was one of the Follow Through models that was *outperformed* by the comparison groups on more than 40 educationally significant differences in reading. Compared to the Direct Instruction model, there were over 70 educationally significant outcomes in reading, all in favor of Direct Instruction. Yet, a study that lacked rigor (insufficient number of subjects, many intervening variables that could account for outcomes) served as the basis for a demur about whether instruction should be highly structured or informal and whether the long term effects were "clear." Punctuating the absurdity of this juxtaposition of the Commission's comments is the fact that there has always been substantial data that if children failed in grade k-3, failure in later grades is virtually guaranteed. Given that High/Scope children failed in k-3, the claims of the preschool study should have been viewed with skepticism.

Direct Instruction is also implicated in connection with the Commission's conclusions about grouping children for reading. In this section, the Commission presents a host of ideas supported only by arguments, not real data. One of the recommendations is that

> . . . reading groups do not always have to be formed on the basis of ability. For example, the advantage of small group instruction for holding attention would still be there if children were sometimes grouped on the basis of interest in the topic. (p. 91)

The assertion seemed plausible to the Commission, but there is no data to support the idea that this recommendation is either manageable or that it produces results as good as those used by highly effective programs. Also there is the fact that beginning readers and corrective readers tend to make a higher percentage of decoding mistakes when they are engaged in an interesting topic because they forget to decode the words carefully (*Corrective Reading, Decoding B Teacher's Guide,* 1978). Furthermore, no study had provided any convincing evidence that grouping by interest is a manageable practice for a highly effective program.

Direct Instruction insists on strict ability grouping for beginning-reading instruction. Later instruction in reading is efficiently taught to entire classes, but classes that are relatively homogeneous with respect to reading skill. The Commission referred to Direct Instruction's endorsement of some whole-class activities in formulating a recommendation that distorts the facts:

> Because of the serious problems inherent in ability grouping, the Commission believes that educators should explore other options for reading instruction. One option is more use of whole class instruction. This seems feasible for aspects of phonics, spelling, study skills, and comprehension. There are programs that recommend whole class teaching some of the time,[22] and they achieve good results, but whether the results are attributable to the use of whole class instruction or other features of the programs is not known. (p. 91)

The Direct Instruction reading program *DISTAR I and II* (Engelmann & Bruner, 1974) is referenced (note 22) as "evidence" for this assertion. The teacher's guide says nothing approximating this plan.

What to Do About the Recommendations

The Commission made quite a few recommendations, including the following:

- Teachers of beginning reading should present well-designed phonics instruction.
- Reading primers should be interesting, comprehensible, and give children opportunities to apply phonics.
- Teachers should devote more time to comprehension instruction.
- Children should spend less time completing workbook and skill sheets.
- Children should spend more time in independent reading.
- Children should spend more time writing.

Some of the recommendations are based on data of successful beginning-reading programs (such as the recommendations about phonics and the design of the text). Some are largely prejudices, such as the recommendation of children spending more time in independent reading. Beginning readers who are just learning to read do not benefit from independent reading (Armbruster et al, 2001).

In addition to broad recommendations, the Commission made more-technical recommendations. Many of these were not data-based. "For example, the sounds of some letters such as *r*, which are especially difficult to produce correctly in isolation, might be introduced best using the implicit approach" (p. 42). The Commission stops short of saying that the implicit approach *is* best, but the implication is clear. The fact is that Direct Instruction is quite successful in teaching the sound for *r* in isolation. Furthermore, the Commission had no data to suggest that the implicit approach is highly effective.

In summary, the Commission's recommendations constitute opinion, hypotheses, statements of conventional wisdom, and supposition, not fact. No known effective program had ever been designed by following the formula proposed by the Commission. Therefore, the recommendations were largely experimental and arrogant distortions of the knowledge the Commission had about effective reading.

Because empirical verification of the recommendations was lacking, the recommendations are not actually scientific, even though they are adorned with scientific trappings, like references and footnotes. In fact, the Commission used legitimate data to create recommendations that are not consistent with the data. If the report were strictly scientific, it would point out that its recommendations are experimental. Further, it would note the need for research that tested the recommendations before consumers took them as guides to creating successful instruction.

The proposed study could have specified a comparison group that uses a known, highly effective approach, such as Direct Instruction. The comparison of this known program with those based on the Commission's recommendations would have yielded specific information about how well the Commission's recommendations translated into excellent instruction.

A less humane alternative would be to show the inadequacies of the Commission's recommendations by following all of them to the letter and purposely creating a program that performed as poorly as whole-language approaches. The results would show the extent to which the Commission's recommendations would have to be discarded or extensively qualified before they generated effective programs.

Although this suggested research is distasteful because children would be subjected to programs that are designed to fail, the concern is a little paradoxical because children are routinely subjected to beginning reading programs that are experimental and generate high rates of failure; yet these placements are endorsed by school districts, school boards, states, and the publishers who create them. The only difference is that these programs are not created to fail. They are simply products of ignorance about what works, particularly with lower performers.

The Commission's recommendations would be far less presumptuous if they had remained within the context of specific programs. The report would not have to address which properties make the program more effective, simply that this is a concrete example of an efficient sequence that has whatever variables are needed to make it efficient in phonics and other components.

The Presumed Consumer

A final issue is exactly who is supposed to use these recommendations? The recommendations are apparently ubiquitous, useful to the publisher and to the teacher. If the recommendations are to be used by the district, state, school, or teacher, however, they tacitly provide a portrait of a consumer that does not really exist. The portrait characterizes teacher and school as creators of instructional material, but this characterization does not fit the data. Schools and teachers are purchasers of published packages, not designers of instruction. This fact seems apparent from the amount of money that is spent on sales of beginning-reading material.

Because teachers and schools are largely consumers of packages and services, not creators, the Commission's report does not address their central information need. As informed consumers, they need to know what they can use to solve their instructional problems—which programs or services lead to highly effective results. They would not be particularly concerned with programs that are slightly better than what they are using now, or with an average-performing program that has a certain set of features, but that failed with lower performers.

A report that does not provide information about currently available programs that have data of superior effectiveness does not answer the primary question the informed consumer would have. The argument most frequently used to explain why specific programs may not be mentioned or endorsed is that it is not permitted. The reason for it being not permitted is that it is somehow unethical. Given that it is not unethical to test cars and rate their performance on safety, or to license drugs that have met performance standards, there is no compelling reason why education cannot present facts involving performance of specific programs, particularly since these programs are clearly named in the studies that generate the information that they are successful.

In summary, neither the designer nor the consumer could create highly effective programs from the type of information *Becoming a Nation of Readers* provides. They could create results that are better than what they are achieving if they use programs that have none of these features. However, these results would not generate highly effective programs because they don't address the litany of detail that is necessary to create a highly effective program. If the goal is to identify and create highly effective programs, however, the recommendations are seriously flawed.

Basic Research Today

Since the publication of *Becoming a Nation of Readers*, there has been extensive delineation of some of the categories identified by the Commission on Reading. For instance, phonemic awareness has become a popular topic. The reporting problems and distortions of fact identified in *Becoming a Nation of Readers* still occur, however. The consumer's questions about which programs have been demonstrated to create highly effective outcomes are not answered and the recommendations continue to assume that the consumer is a program creator.

For example, the 2000 *Report of the National Reading Panel: Teaching Children to Read* is replete with traditional wisdom and dangerous assumptions, even though the report is based on a higher quality of data and the Panel is more articulate in identifying the features of exceptional programs. The report, however, has the same basic problem of assuming that a common feature of successful programs is both necessary and sufficient to use as guidelines for creating highly effective programs. Furthermore, the report provides a greatly distorted view of the consumer and the consumer's capacities. The report may actually present more opinion as fact than *Becoming a Nation of Readers* did.

Phonemic Awareness

The Panel's comments on phonemic awareness suggest that the teacher is something of a constructor who has the instincts and the dedication needed to conduct a systematic scientific inquiry necessary to learn about the relative attributes of different approaches to phonemic awareness (PA).

> First, PA training does not constitute a complete reading program. Rather, it provides children with essential foundational knowledge in the alphabetic system. It is one necessary instructional component within a complete and integrated reading program. . . . Second, there are many ways to teach PA effectively. In implementing PA instruction, teachers need to evaluate the methods they use against measured success in their own students. Third, the motivation of both students and their teachers is a critical ingredient of success. Research has not specifically focused on this. (p. 8)

The Panel presents an experiment as if it has a basis in data. If the teacher understands that PA is only one of the necessary components of a "complete and integrated reading program," the implication is that the teacher will need to combine PA with the other components. Where is the data that teachers are able to do this successfully? What evidence suggests that the teacher has the time, the knowledge or desire to engage in this enterprise? And how do the results compare with a prepackaged program, such as a Direct Instruction reading program? This program would make a good baseline program for comparing the attempts of teachers to construct their own package by combining components.

The Panel's suggestions about how teachers should "evaluate" the success of their selected or created approach is glib. How many years would it take for the teacher to "discover" or "create" an excellent combination (given that it would be hard to try out more than one or two combinations a year)? What kinds of records would be needed to make this enterprise systematic? And how does this pursuit fit in with the reality of the classroom, which is that the district has adopted some material and the teacher is expected to use it? There are probably only limited funds for the teacher to conduct her ongoing analysis. Most relevant, both history and demography suggest that it is unlikely for the teacher to develop a highly effective package.

The Panel's observation about motivation of teachers and students being critical to success is curious. The facts of performance show that if teachers behave in certain ways, the likelihood of their students performing better is greatly increased. Given these facts, what is the purpose of discussing motivation, particularly if the Panel has no data on its effects? If "motivation" is to become a factor or variable of a successful program, it must first be operationalized and expressed as things the teacher does to demonstrate and instill "motivation."

The Panel does not identify exemplary programs that have PA packaged with the other ingredients of an effective program. Yet, the Panel has information that there are programs that are successful. If they are successful, wouldn't it follow that whatever phonemic awareness they convey to the children is adequate?

Phonics

The Panel's treatment of phonics has the same problems as its treatment of PA. No specific exemplary programs are identified; the description is limited to features, which are liberally interpreted by the Panel; the needs of the consumer are not realistically addressed.

The Panel drew this general conclusion from the meta-analysis:

> . . . systematic phonics instruction enhances children's success in learning to read and that systematic phonics instruction is significantly more effective than instruction that teaches little or no phonics . . . A variety of systematic phonics programs have proven effective with children of different ages, abilities, and socioeconomic backgrounds. (pp. 9–10)

The last sentence is absolutely unsubstantiated by any data. There is not a variety, but very few. Further, it is misleading to suggest that some work with low performers but not with higher performers. This relationship has never been demonstrated. If a program works well with the lowest performers it will be effective with all children. The only difference is the rate at which students progress through the program sequence. For higher performers, the teacher should be able to proceed at possibly three or four times the rate required to bring very low performers to mastery. Project Follow Through provided some evidence of this relationship. In each site there was a "mix" of middle-class children. The Direct instruction model had by far the highest performance for this subgroup.

The Panel also drew some conclusions that suggest non-data-based prejudices:

> Some phonics programs showing large effect sizes require teachers to follow a set of specific instructions provided by the publisher; while this may standardize the instructional sequence, it also may reduce teacher interest and motivation. Thus, one concern is how to maintain consistency of instruction while still encouraging the unique contributions of the teachers. (p. 10)

The Panel's reasoning is curious, if not contradictory. Programs that show "large effect sizes," by definition work far better than average programs with students from comparable populations. If large effect sizes are created by unmotivated teachers, the issue of motivation is not very important. If "motivation" is necessary, however, it would be impossible for the large effect sizes to come from any approach that had unmotivated teachers. In other words, the evidence of effectiveness would suggest that in addition to standardizing, the programs—when properly implemented—assure that teachers are involved and are performing well.

The Panel may not understand how a program works, but the Panel's conclusion about the need to maintain the teachers' unique contributions implies that the teacher should not follow the program too closely. Again, there is not a shred of data to suggest that when teachers tinker with the sequence or provide enrichments that are outside the specifications for presenting an effective program, student performance will improve. For Direct Instruction, there is a strong positive correlation between the extent to which teachers follow the program and the performance of their students.

The Panel observes,

> [It is] important that teachers be provided with evidence-based preservice training and ongoing inservice training to select (or develop) and implement the most appropriate phonics instruction effectively. (p. 11)

The Panel's reference to evidence-based training is puzzling. The notion that programs that have certain features will be effective is not evidence-based. The interpretations the Panel provides are not evidence-based. The reference to selecting, developing, and implementing the most appropriate phonics instruction effectively implies that it is possible for teachers who have never used highly effective instruction will be able to learn about it or how to develop it from inservice and preservice training. This possibility is not evidence-based. Inservices that deal with principles or discussions as general as those provided by the Panel have no data of effectiveness.

The most relevant fact about effective inservice and preservice is that it must be program-specific. Program-specific inservice for DI is highly correlated with improved student performance.

In other words, teachers are not expected to invent the programs, but need to learn about using material that has already been adopted and that has the potential to achieve "large effects." The more the training time deviates from this goal, the weaker the goal becomes.

The Panel observes,

> Teachers should be able to assess the needs of the individual students and tailor instruction to meet specific needs. However, it is more common for phonics programs to present a fixed sequence of lessons scheduled from the beginning to the end of the school year. In light of this, teachers need to be flexible in their phonics instruction in order to adapt it to individual student needs. (p. 11)

This may be the most distorted recommendation the Panel made. It is self-contradictory, illogical, guilty of creating a false dilemma, and completely without substantiating empirical evidence. Simply because a program has a fixed sequence does not imply either that all students begin at Lesson one, that they progress through the program at the same rate, or that they complete exactly one level of the program during the school year. The Panel has apparently confused practices for meeting the individual needs of students with the structure of the program. The Panel's confusion is that it has the mistaken belief that all students go through the program at the same rate. No, this is where individual differences come into play. Students who are placed in an effective sequence do not go through the program at the same rate. The goal is for the teacher to teach each lesson to mastery. Some students need more practice to achieve mastery than others. That's why students should be homogeneously grouped. All students in a properly configured group require about the same amount of practice, so the teacher is able to teach to mastery at a pace that is appropriate for all. In other words, the individualization is not a variable in *what* students learn, *how* they learn, the sequence in which it is presented, or the amount of material that constitutes a lesson. It's simply the amount of practice students need to achieve mastery.

The first step in teaching to mastery is to place students appropriately in the sequence. If the program is designed so that each lesson teaches a small percentage of new material and reviews everything introduced in the preceding few lessons, the program is capable of accommodating all students who lack skills that the program teaches. The teacher simply finds the lesson at which the student is around 75-80% correct on material the lesson presents. The student will be able to achieve mastery on this lesson and on all subsequent lessons.

The recommendation that teachers should not to follow the program strictly, but to tailor instruction, and to be flexible, implies that the teacher knows how to do this. Virtually all of the teachers who have never used highly effective programs have no idea about how to create highly effective instruction. Again, there is data that if teachers teach DI sequences effectively, their children will progress. There will be virtually no non readers.

Need for Research on Basic Assumptions

The final excerpt reveals the Panel's assumptions about the teacher's capabilities:

> [It is] important to note that fluent and automatic application of phonics skills to text is another critical skill that must be taught and learned to maximize oral reading and reading comprehension. This issue again underscores the need for teachers to understand that while phonics skills are necessary in order to learn to read, they are not sufficient in their own right. Phonics skills must be integrated with the development of phonemic awareness, fluency, and text reading comprehension skills. (p. 11)

The Panel issues these cautions but does not provide any consumer information about what the teacher should do to heed them or which, if any, specific programs provide what the Panel considers an adequate integration of these components. The assumption is that teachers will somehow be able to use the information about what to integrate to create effective programs.

This hypothesis should possibly be researched —the degree of sophistication of the consumers, the effectiveness of their program adjustments or creations, and the usefulness of the information the Panel provides to this consumer. In schools that are currently failing, the teachers have failed. If we use these demographic facts to infer the skill level of the overwhelming number of k or first-grade teachers, we don't see sensitive diagnosticians. We see people who have never been trained in even the basics of effective instruction, people who are effete and clumsy at meeting the individual needs of the children. After all, a very small percentage of the children perform at or above the 50th percentile.

For these teachers to use their experience and the Panel's recommendations to discover or formulate highly effective reading instruction, they would have to solve four problems:

> 1. *The problem of program development and evaluation.* The Panel provides the formula that the teachers should be able to try different approaches, evaluate them, and arrive at an effective solution. This involves using information about formative outcomes to infer what should be done. Given that the teachers have never been able to do this in the past, it seems presumptuous to assume that they will know how to do it now. So some training would be required in using information, identifying efficient ways of mixing and matching components, and of using performance data to make efficient changes in strategies (whether to slightly modify a component or scrap it and start over).
>
> 2. *The problem of dead reckoning.* Dead reckoning is navigating without sufficient feedback, which means navigation miscalculations may go unnoticed. The teachers who try to follow the *P*anel's formula would have this problem. Even if the teachers had guidelines and had learned program-design rules that would permit them to proceed efficiently, they would have no benchmark for comparing their attempts with the results attainable with a highly effective program. Let's say that the teachers formulate a combination that does better

than the current program, but let's say that the level of student performance is far below what is possible with an excellent program. How do the teachers know that they are not even close to having an excellent program? From their standpoint, the new sequence appears to be wonderful. They see children more engaged; they see a higher percentage of children learning. Without some comparative information about what is possible with an excellent program, they would understandably conclude that they had achieved excellent results. If the teachers had no information about performance levels of students in effective schools, many teachers and schools in the Panel treatment would be shocked to discover later how far below the level of the effective program they performed.

3. *The problem of using programs that are less than highly effective during the years of program development.* This is an ethical problem. Even if the teachers were able to design or select an effective combination of components, it most probably would not happen in less than three years. During the three years of "discovery," the children served by these teachers would be subjected to instruction that was far less effective than would have been possible if they had been in a highly effective sequence. Because the probability is low that the teachers would have discovered or created a highly effective sequence even after three years, the Panel's recommendations for experimentation seem to have a serious ethical problem, possibly something like the one involving forceps that saved lives of babies during difficult deliveries. For years, forceps were a trade secret for a small group of physicians. If the *P*anel knows that there are highly effective programs and does not name them but instead tells teachers to figure out their own combination, the decision seems to have serious problems, particularly because the likelihood of teachers being successful is neglig*i*ble. Recommendations that probably will result in poor instruction during (and possibly beyond) the formative years seem unjustified on more than a small scale. Furthermore, a small-scale experiment would be justified only to show the difference between what teachers could create and highly effective programs like DI*.*

4. *The problem of dissemination and cost.* If even as many as one out of five teachers developed effective programs, the cost of the experimentation with different components would be at least three times more expensive than that of installing a highly effective program now. Furthermore, this cost of installing or disseminating the effective program is not eliminated, simply deferred until some teacher perfects what seems to be a highly effective package. The problem is that the dissemination would then be pretty much the same as it would be with an effective program that had been available at the time the experiment was undertaken. Other teachers would have to be trained in how to use the program, and so forth.

Teachers are just like students in the sense that some of them have more skills and know more about teaching and managing than others. One of the problems of at least some highly effective programs that teachers may create is that they require an impractical amount of teacher-training, and not all teachers learn what they need to know from even elaborate training. The Panel recognizes this problem, although it does not clearly express the implications.

> Other programs require a sophisticated knowledge of spelling, structural linguistics, or word etymology. In view of the evidence showing the effectiveness of systematic phonics instruction, it is important to ensure that the issue of how best to prepare teachers to carry out this teaching effectively and creatively is given high priority. (p. 10)

Because at least some of the programs that teachers develop will imply extensive teacher training, the programs would fail, not because they are ineffective, but because they are impractical.

An Unbalanced Approach to Training

If the Panel's knowledge is compared with its recommendations, it becomes apparent that the Panel's approach to reporting what it believes to be the truth is unbalanced, and prejudiced by irrational considerations. The Panel knows which programs have achieved excellent results. These are the programs that have generated the information about effective programs. The Panel does not share this knowledge or even offer teachers a significant choice of how to proceed, and the risks of taking different paths.

A "balanced approach" would do something like the following:

1. It would present recommendations providing the option of using an extant, highly effective program or dabbling and developing their own approach. It would list the programs that share or that have generated the information about the features shared by highly effective programs.

2. It would explain the probability of success for each option. The Panel would explain that if one adopts a prepackaged program that has achieved large effects in student improvement, teachers would have to be trained to use the program and follow it closely. The intervention would probably result in a large effect. Large differences in performance would be observed during the first year of the implementation. Teachers would be able to help the children they have not.

If they mix, match, and experiment, the chances of creating a highly successful package in less than three or four years is probably 1 out of 1000, which means that the students being served for the next three years will probably not receive highly effective instruction. If the school does not expect all teachers to invent successful packages, teachers will have to be trained in using effective packages after they have been developed. So there is really no savings in what has to occur, merely when it will happen, the probability of it ever happening, and when students will be served by effective programs.

Perspective on Categories of Content

The greatest problem with the current view of “what works” is that it is limited to what researchers have observed. However, they have not observed very much. There is a lot more to reading instruction than the categories that are currently popular—phonemic awareness, phonics, text decoding, and comprehension. Consider phonics. It is treated as something of a panacea, but how far could phonics take the learner? The idea is that one letter (or sound combination) makes one sound. This works for “Nan had a bad cat,” but it starts falling apart if “Nan had a bad day,” or if “Nan and the other children saw a show,” and still farther if “Nan and a friend were walking to school.” The letter *o* makes seven different sounds in common words. How does the program solve the problem of carefully teaching words with all these sound variations and teaching all the common irregular words? Solving this problem is not only far more challenging than introducing “*a* is for apple” but it must be addressed thoroughly and early in the first level of the program.

Of course there are cavalier ways of avoiding the problems of making everything teachable, discriminable, and supported by adequate practice. The most common is to present some phonics activities and then literally throw irregulars into stories without sufficient practice. For instance, early in Open Court, seven new irregulars are introduced in a single lesson. Although this rate of introduction may be adequate for a few children, it most certainly is outrageous for typical at-risk populations.

Until researchers go beyond superficial observations of a program and gain some understanding of what is needed to make teaching manageable for the teacher and the students, panels will continue to believe in half-truths and make recommendations that are naïve.

Impediments to Creating Sensible Programs

The largest impediment against the commercial development of superior programs is the untenable relationship between states that have statewide adoptions and publishers. During a cycle of possibly seven years, the state formulates adoption criteria for the next submission and issues these to publishers. By the time the details are in place, the publisher interested in submitting a program has possibly a little over a year to design a program that ostensibly meets these criteria. There is insufficient time for the publisher to field-test the product and evaluate whether its interpretation of the adoption criteria result in excellent performance. Instead, the product is assembled so that it gives the appearance of having all the features the criteria require. The state evaluators are not people who have a record of being highly successful. They are just teachers, administrators, and possibly parents. They inspect the submission, comparing the features with the features described by the criteria.

The result is raw, poorly sequenced programs that are field-tested for the first time when teachers and children first use the material. An analogy would be a car that had never been field-tested before being put on the market, or a drug that had never been evaluated on any human subjects (or possibly any subjects).

In addition to the time problem, the relationship between publisher and adopting agencies is sick for four reasons:

1. It is based on the assumption that the psychology and knowledge of entering kindergarteners or third graders changes so much in seven years that a new program with new goals and objectives is needed. In fact, the need for a new

program is created by the failure of the last set of adopted programs. Also, the same programs that were effective in 1968 are effective today, but states and districts apparently have not figured this out yet.

2. The relationship assumes that the failed district or failed state has sufficient knowledge to specify how an effective program is to be designed. Consider a district like New York City, Chicago, Denver, or any of the other failed districts. The administration in these districts knows precious little about what it takes to provide effective instruction for at-risk students. Yet, these districts arrogantly assert that each of their frameworks or criteria for new programs will work. They design procedures to assure that publishers comply with the districts' latest prejudices.

3. The districts and states spend billions on adopting programs but never see fit to first field-test them on a small scale and determine the extent to which they are effective in the classroom, particularly with at-risk students. With field test results, the district would at least know that it if installed certain programs, extensive failure would result.

4. The relationship victimizes the school, the teachers, and the students. The school enters this relationship with concerns about adopting effective instruction. The guidelines or framework assures the school that the adopted programs are "based on the most recent research on how children learn." In the end, the adopted programs fail for one basic reason: they are poorly designed. The fact that DI programs were effective but have been overwhelmingly rejected by states and districts because they fail to meet the standards of states and districts implies that the standards and criteria did not provide schools and teachers with adequate information about what works.

Legacy of Basic Research

Basic research in education attempts to mimic the form of basic research in the physical sciences, where the goal was to discover general laws; however, there is no close parallel between instruction and elementary laws of a science like physics. For bodies in motion, rate, direction, friction and gravity are about the only major variables that complicate calculations. For instruction in something like teaching reading there are many variables that affect the outcome.

The inputs that cause changes in behavior are grossly different for teaching something and for altering the motion of an object. The extent of the difference is demonstrated by the fact that we can start with two identical objects when we measure motion. There are no two learners that are identical. For bodies in motion, the motion is either directly measured or inferred from a relatively small number of principles. For learning, the inferences are elaborate. The actual reading behavior is so remote from observation that it is possible to test any reading discrimination or skill by using *any* pair of responses in the learner's repertoire. For example, "Touch the table if I read the sentence correctly. Point to the book if I make a mistake…". We see the behaviors of touching the table and pointing to the book. We infer reading or reading problems.

To understand the science of effective instruction is to understand the lawful aspects of the learner, the rules for communicating any information effectively to a naïve learner, and both the content and principles for dissecting it, simplifying it, and parceling it to the learner a bit at a time. An adequate set of principles could not be based on less information.

If the goal of basic research is to provide information about how to teach something effectively, the model must be changed to one that recognizes the need for *concrete examples that show control of particular variables within the context of the other variables.* The consumer needs to understand the details of one or more successful scheme, not slogans about how success is generated. With this concrete orientation, the line between basic research and applied research disappears. The research would tend to follow a Type 2 approach format. The discussions of products would identify features within the context of specific programs— not global abstractions and not concrete examples that lack empirical validation.

Just as an automobile report may compare the carburetors of two highly successful cars, educational research studies could discuss the details of the PA components of two highly successful programs. The report might stray into conservative speculation, but the kind of speculation that characterizes the Panel's report would be proscribed. The report would indicate precisely how the PA component is linked to the beginning-word-decoding activities and how economies in correcting mistakes are achieved through the PA component.

The myth of general principles like teaching phonemic awareness would be replaced by information about how it may be efficiently engineered in a way that anticipates how the learner will later use PA skills to decode words. This step in specificity is a step toward a more realistic, scientific understanding of instruction.

References

Anderson, R. C., Hiebert, E. H., Scott, J. A., Wilkinson, I. A. G. with contributions from members of the Commission on Reading (1984). *Becoming a nation of readers: The report of the Commission on Reading.* Washington, DC: The National Academy of Education, The National Institute of Education, The Center for the Study of Reading.

Armbruster, B. B., Lehr, F., & Osborn, J. (2001). *Put reading first: The research building blocks for teaching children to read.* Washington, DC: U.S. Department of Education.

Becker, W. C., & Engelmann, S. (1978). *Analysis of achievement data on six cohorts of low-income children from 20 school districts in the University of Oregon Direct Instruction Follow Through Model.* (Follow Through Project, Technical Report No. 78-1). Eugene, OR: University of Oregon.

Engelmann, S. (1992) *War against the schools' academic child abuse*. Portland, OR: Halcyon House.

Engelmann, S., et al. (1978). *Corrective Reading, Decoding B Teacher's Guide.* Chicago: Science Research Associates.

Engelmann, S., & Bruner, E. C. (1974) *DISTAR I, II.* Chicago: Science Research Associates.

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.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. Wittrock (Ed.), *Handbook on Research in Teaching* (pp. 376–391). New York: Macmillan Publishing Company.

Schweinhart, L. J., & Weikart, D. P. (1980). *Young children grow up: The effects of the Perry Preschool on youth through age 15.* Ypsilanti, MI: High/Scope Educational Research Foundation.