P. David Pearson
Michigan State University, East Lansing, USA

The First-Grade Studies: A personal reflection

It is a special privilege for me to comment on the First-Grade Studies during our commemoration of the 30th anniversary of their publication. They have a special place in my memory and my personal history, for it was during the early years of my graduate study at the University of Minnesota that Bond and Dykstra were completing the final stages of data analysis and manuscript development for the Cooperative Research Program in First-Grade Reading.

Remembrances of things past

Several images of those times and the people associated with the Studies leap readily to mind.

• Most vivid is the image of my fellow graduate student, John Litcher (now in the School of Education at Wake Forest University), trudging across the wintry Minnesota landscape with, quite literally, boxes of punch cards to feed into our state-of-the-art CDC computer. That computer, which occupied a space the size of two ordinary college classrooms, served the computing needs of the entire research community on campus with roughly the computing power of the typical 486 machine that now sits atop the computer table in our offices. Somehow, we managed, but not without many trips across that wintry landscape.

• Another image—the day the first batch of final reports (on which the 1967 *RRQ* article was based) arrived at Bob Dykstra's office. What was so special was the sense of excitement that we all felt as graduate students— as if we were in possession of privileged information, as if we had the inside scoop on a major professional secret. And we did, for a few days at least. It was also the first occasion on which I met Guy Bond (who, I learned later, had reviewed my application to graduate school).

• Graduate seminars in which we examined the First-Grade Studies in juxtaposition with Jeanne Chall's newly released book, *Learning to Read: The Great Debate* (1967), and some of the interpretive pieces written in the wake of the momentous report. Of particular interest among us was the question of whether the Studies did show that code-emphasis approaches were superior to meaning-emphasis approaches in promoting reading achievement.

• Frequent discussions with John Manning, my advisor and a principal investigator for one of the 27 individual first-grade studies, about the nature of beginning reading instruction, the early reading program that he had put together in Clovis, California, and how the statistical practice of analysis of covariance had obscured the positive impact of his kindergarten intervention program on first-grade reading achievement. You see, a major component of John's intervention was an intensive readiness program in kindergarten in which kids learned letter names and sounds as a part of a larger program to promote attention and persistence, the assumption being that if students can acquire these dispositions, they can learn in large classrooms in which group instruction requires vicarious learning. Of course, this treatment effect for students in his experimental group was wiped out when the analysis of covariance was conducted. The data on p. 65 of the original text/p. 380 in this volume corroborate John's concern.

Two different worlds

Enough of reminiscence. My task in this retrospective is to examine the First-Grade Studies from the lens of literacy research at the *fin de siècle*, and to ask what they can teach us today. To examine what they might

teach us, we must first acknowledge just how different our world of reading research is now from the world in which Guy Bond and Robert Dykstra analyzed the data and wrote the text for the First-Grade Studies. We were on the cusp of several revolutions. Dick Venezky had just shown us that English orthography, when examined from a nuanced linguistic (morphophonemic) perspective, was uncommonly predictable. Ken Goodman was putting the final editorial touches on *Linguistic Cues and Miscues in Reading* as he began his quest to convince us all that reading is fundamentally a language process. Bob Ruddell and Harry Singer were probably planning the 1968 institute that would lead to the first publication of *Theoretical Models and Processes of Reading*. Psycholinguistics was emerging as a field, as was sociolinguistics, through the work of scholars such as William Labov, Joan Baratz, and Roger Shuy, who were using it as a lens for examining the role of dialect in learning to read. We were still nearly a decade away from the cognitive revolution. Postmodernism, constructivism, feminism, and critical theory, while surely alive and well in well-tended intellectual plots, were still almost 2 decades away from exerting a major influence on mainstream thinking about reading processes or practices.

Despite these rumblings from our sibling disciplines, it was fair to conclude, in 1967, that researchers and practitioners viewed reading as fundamentally a perceptual process. Whether we supported phonics, linguistic approaches, initial teaching alphabet (i.t.a.), or look-say as the most appropriate approach to beginning reading, the common view of most (by no means all) educators was that the job in reading was to turn symbols into sounds in order to get the words right and, in so doing, receive the meaning set down by the author. I belabor this point about theoretical contexts and trends so that modern readers of this remarkable inquiry can judge its worth as an intellectual endeavor from the perspective of the theoretical constructs that prevailed in the late 1960s.

The legacy of the First-Grade Studies

So what can we learn from this 30-year-old document? How does it speak to issues, concerns, and questions faced by the current generation of reading educators?

Learning about the efficacy of early reading instruction

If the only document you read is the 1967 *RRQ* version of the First-Grade Studies, the most plausible conclusion about the most effective approach to teaching beginning reading is, "It all depends." Across sites and

projects, the variability is remarkable; as Bond and Dykstra conclude, "Reading programs are not equally effective in all situations." Statistically, this means that there were a fair number of project-by-treatment interactions: A treatment that rose to the top in one project (project equals site) may have achieved poor results in a second.

The second conclusion that you (as well as Bond and Dykstra) would draw is that pretty much any alternative (basal plus phonics, phonics-first, linguistic, special orthography, or language experience) was, on balance, superior to the whipping boy of the First-Grade Studies, the conventional look-say basals popular in the early 1960s. In fact, Bond and Dykstra use this finding to support two interesting conclusions: (a) that combination approaches are superior to single approaches, and (b) that reading instruction is amenable to improvement (apparently on the assumption that basals represent the conventional wisdom that stands in need of improvement).

You would not conclude from the 1967 *RRQ* report that code-emphasis approaches were superior to meaning-emphasis approaches. Even though both the phonics plus basal and the phonics first (dubbed phonic-linguistic) approaches consistently elicited higher comprehension and word reading scores than basal approaches, Bond and Dykstra avoided the conclusion. In a later article in which he compared Chall's (1967) conclusions with the span of first- and second-grade results from the First-Grade Studies, Dykstra (1968a) concluded, "Data from the Cooperative Research Program in First-Grade Reading Instruction tend to support Chall's conclusion that code-emphasis programs produce better overall primary grade reading and spelling achievement than meaning-emphasis programs" (p. 21).

One of the most consistent findings in the Studies (also replicated in the second-grade follow-up) is the singular absence of aptitude-by-treatment interactions. With the exception of the Language Experience analysis (in which, depending on the site, LEA proved uniquely suitable to either high- or low-aptitude students), there were no indications that methods were consistently effective with subgroups of students identifiable by various aptitude indices (e.g., intelligence, letter-name knowledge, or phoneme knowledge). The commonsense homily about low-aptitude students needing the structure and guidance of an explicit phonics-first approach received no support from any of these analyses. Where phonics worked well, it worked well for all students; conversely, when it worked poorly, it worked poorly for all.

Learning about contextual variables

As a part of the data-collection process, information was gathered about teacher and classroom variables (teacher experience, class size, etc.). While few of these

variables showed much relationship to student achievement (they tended to correlate in the .2 to .4 range), another contextual variable, project, emerged as a powerful factor in explaining student differences. While project was mentioned in the reports of both the first- and second-grade analyses, it was extensively analyzed only in the full version of the second-grade extension of the studies, in which Dykstra (1967) compared the means for each instance of a method across projects. A common finding was that the mean for the lowest performing method in Project X was often as high as or even higher than the highest performing method in Project Y. Even controlling for a wide range of individual difference variables, there was apparently something about a particular project that elicited high (or low) performance for all students regardless of method. This sort of finding led Dykstra to conclude that future research needed to focus on site-based variation in order to learn what it was about the *culture* of a site that led to such consistently exceptional performance. His recommendation presages our current preoccupation for examining instruction and learning with a *situated* lens.

Learning about the practice of research

There is a great deal to learn about the practical concerns of conducting research, especially in settings in which many individuals are involved.

1. *Other things being equal, include multiple measures of any important phenomenon.* Clearly the principal investigators of the 27 individual studies thought it important to include multiple measures, although they have many more measures of word reading than of comprehension. It should also be mentioned that writing samples were included in the original design, but were not included in the final analysis (apparently out of concern for standardization of administration and concerns about scoring reliability). The exclusion of writing samples was a real disadvantage for the i.t.a. treatments. One of the serendipitous findings of some of the specific i.t.a. analysis was that when kids were equipped with a transparent orthography that was completely under their control, they became fearless writers, producing a great deal of text. One is reminded of the remarkable fluency of students in today's classrooms when they are encouraged to use invented spellings (e.g., Clarke, 1989).

2. *Sometimes finding an appropriate control group requires a bit of imagination.* Given the variety of interests brought to the table by the 27 principal investigators, it is truly amazing that they were able to settle on any common notion of a control group. The idea of using a *garden variety* basal, which according to research of the time was used by 95% of all schools in the United States, was a brilliant compromise. Granted, it did not al-

low for comparisons among the various innovative procedures, but it clearly allowed for a common benchmark and for some interesting cross-site analysis. The premise that all basals are created equal (and therefore constitute a common basis for comparison) does require a stretch but, given what we know of basal production at the time, it is probably not a far-fetched assumption.

3. *The appropriate unit for statistical analysis depends upon the question one wants to answer.* For the correlational analysis, Bond and Dykstra used the scores of individual students as the basic unit of analysis, a practice that fits the predictive nature of the question (predictions are about individuals, not classes). But for the comparison of methods and the method by treatment interaction analyses, class means (actually the separate means for boys and girls—so that gender could be included in the analysis) were used. Again, they fit the question and the situation; the method was applied simultaneously to all members of a class.

4. *Novelty can affect findings.* Both the classic Hawthorne effect (all the groups do better because they are in a study) and the more selective novelty effect (the experimental groups do better because they get privileged treatment in comparison to controls) can compromise findings in experimental studies. While attempts were made to control for novelty (e.g., making sure that even the lowly basal teachers received an equal amount of inservice training), it is hard to imagine the same feeling of newness and excitement among teachers who are part of a business-as-usual treatment. While Bond and Dykstra did not mention this possibility, other commentators on First-Grade Studies did (e.g., Sipay, 1968; see also Southgate, 1966).

5. *Treatment fidelity is a common problem in large-scale research.* In the limitations section of the second-grade extension (1968b), Dykstra explicitly mentioned the problem of treatment fidelity across sites: What was called a linguistic approach in Project X might be quite different from a linguistic approach in Project Y. (Curiously, there is no limitations section in the *RRQ* version of the First-Grade summary.) Equally problematic was ensuring the fidelity of treatments within projects. While every attempt was made to ensure fidelity, teachers occasionally applied their own standards to a method in order to make it fit their own philosophy and professional practice. (Dykstra [personal communication, 1968] tells a story about visiting a remote Language Experience site as a part of his monitoring role and finding the teacher engaged in a lesson from a commercial phonics workbook. When queried about the practice, she replied, "Well, this Language stuff is very interesting and the students really like it, but, you know, they need their phonics...." I am reminded of the incredible variation in what goes on in the name of whole

language in today's instructional milieu. I am also reminded of the highly eclectic tendencies of the teachers nominated as outstanding in Pressley's 1996 work on exemplary teachers.) Of course, the larger the study, the greater the threats to fidelity. (An alternative argument is that if effects survive the probable treatment infidelity that is likely to occur in large-scale instructional work, then they must be truly robust!)

6. *Occasionally technical standards must be compromised for the sake of credibility and utility.* In several situations, such compromises can be seen in the First-Grade Studies, but they are not bothersome. To the contrary, they permit more careful and considered inspection of findings across the different analyses. For example, in the guidelines for analysis (pp. 47–49/p. 369), a logic is put forward that IF and ONLY IF project-by-treatment interactions emerge will separate analyses be conducted for treatments within each of the projects. In truth, the within-projects analysis was conducted regardless of whether project-by-treatment interactions appeared. In another example, parsimony would dictate that a single model for analysis of covariance be selected (e.g., either all of the covariates or only those most likely to control for relevant pre-experimental variation). As it turned out, both a full and a minimal set of covariates were used in all of the analyses. For the reader who wishes to make some of his or her own eyeball comparisons, it is most useful to have the full set of analyses, quite irrespective of whether they are technically appropriate.

Learning some lessons to guide our future research

A common, but usually implicit, standard for evaluating the legacy of a piece of research is whether it generates additional studies on the issue, topic, or question. By that standard, the First-Grade Studies were a dismal failure, for they (in conjunction with Chall's book) marked the end of methodological comparisons in research on beginning reading (at least until the 1990s). Dykstra (1968b) recognized and championed this uncommon legacy:

> One of the most important implications of this study is that future research should center on teacher and learning situation characteristics rather than method and materials. The extensive range of classrooms within any given method points out the importance of elements in the learning situation over and above the materials employed.... The elements of the learning situation attributable to teachers, classrooms, schools, and school systems are obviously extremely important. Reading instruction is more likely to improve as a result of improved selection and training of teachers, improved in-service training programs, and improved school learning climates, rather than from minor changes in instructional materials. (p. 66)

By the way, in the aggregate, the principal investigators (see the Appendix of the reprint of the First-Grade Studies in this volume) involved in the larger set of First-Grade Studies understood the need to move beyond the racehorse mentality. Chall, for example, conducted an intensive study of the instructional practices used by committed code-emphasis and meaning-emphasis teachers (foreshadowing our current emphases on understanding effective practice). Heileman studied the impact of alternative models of inservice training (also foreshadowing our modern commitment to the idea that the pay-off for teacher learning is student learning), and Morrill studied the impact of one-on-one versus group (learning community?) approaches to teacher supervision. And Horn anticipated our current debates over bilingual programs by comparing intensive aural-oral programs in English versus Spanish for students speaking Spanish as a first language.

As it turned out, Bond and Dykstra were prophetic in suggesting an end to methods research. By the early 1970s, we had declared a moratorium on racehorse studies, we had begun the process of changing basals in keeping with the conclusions reached by Jeanne Chall (1967) (and supported, at least in the minds of many, by the First-Grade Studies), and we turned our intellectual attention to new issues, new perspectives, and new disciplines—psycholinguistics, cognitive science, theories and models of the reading process—that lay in wait just beyond the horizon, poised to capture our hearts and minds as a profession. Only recently, driven by politics and alarmist interpretations of test scores and fueled by new sources of funding from outside the educational research industry (Lyon & Chhabra, 1996), have we returned to the question of the best method of teaching beginning reading. The fact that the most recent request for proposals for a national reading center requires an emphasis on early reading indicates that we have pretty much come full circle, back to the issues and questions that prompted us as a profession to undertake the First-Grade Studies some 35 years ago. But, as I indicated at the outset, we are at a very different time and place than we were when all of this began. We have new tools and new lenses for asking and answering questions of teaching and learning.

Even though our world on the cusp of the 21st century is very different, I do hope that we can recover one of the most endearing and important qualities of the era—the model that guided the cooperative part of the endeavor. I think it is a good model for literacy research in a postmodern world.

Like most models, it has a historical antecedent. When I first encountered the First-Grade Studies, I was immediately reminded of a metaphor stolen from that

decidedly premodern thinker, the renown British roman-
ticist, philosopher, designer, and social reformer of the
late Victorian era, William Morris. As a metaphor for his
ideal society, Morris chose the Gothic cathedral; as a
contrastive metaphor for his evil society, the neoclassicist
cathedral.

What Morris liked about the Gothic cathedral had
more to do with the process of construction than the
outcome, although the two are linked. To him, it repre-
sented the proper balance of order and individual free-
dom. There was a master planner with a master plan, to
be sure, but that person was more of a foreman than an
architect. Each worker was assigned responsibility for
completing a particular section of the cathedral. There
was enough coordination between sections to ensure the
structural integrity of the edifice, but no more. Within in-
dividual sections, each worker exercised a great deal of
individual prerogative. As a result, Gothic cathedrals lack
the unity and precision of their neoclassicist counter-
parts. The design of the gargoyles in one corner is quite
different from those in another. The carvings, even the
stained glass, in one section may or may not match
those in another. In every nook and cranny, there is the
distinct mark of the individual craftsman. There was, to
use modern terms, ownership and empowerment.

By contrast, the unity and precision of the neoclas-
sicist design was mirrored in its construction. Workers
carried out orders. They implemented the plans of oth-
ers. There was no room for the individual signature of
each worker. There were no sections individual workers
to point to as theirs. So, said Morris, and partially in re-
sponse to growing Marxist sentiment in England and
Europe, let's build a society of Gothic cathedrals as a
way of ensuring that human beings are connected to,
rather than alienated from, their work.

I like that metaphor a great deal for thinking about
our research, both in its micro- and macroscopic aspects.
Within a project, when a group of us work together—
teachers, administrators, and researchers—we need to
find ways for each of us, as individuals, to put our own
personal stamp on the project. There must be room for
variation. But just as surely as there is variation, there
must also be theme, a common core to which we are all
committed. We should all be learning something differ-
ent about the same thing. We can move that model one
level up and think about implementing cross-site studies,
studies in which teams decide to pool their intellectual

and material resources to gain variable insights on ques-
tions of common interest.

This is, I believe, the model and the metaphor that
guided, knowingly or unknowingly, the First-Grade
Studies. Twenty-seven individual researchers or groups,
each with his, her, or their own agenda, had a unique
piece of the Gothic cathedral of reading research to
shape in a unique image. Each, however, ceded some
independence to be a part of a larger effort, to answer
some bigger questions, than a single study could answer.
I suppose that this sort of collaboration came with the
territory so to speak, since the First-Grade Studies were
funded by the Cooperative Research Branch of the then
United States Office of Education.

It is a model, however, that I think we should
strive to emulate, to reincarnate, both in spirit and in
form. We need both theme and variation in our work.
The themes bring us together, encourage us to share and
collaborate on a common vision, while the variations re-
mind us to respect, enjoy, take pride in, and, most im-
portant, learn from our differences. And to think that our
intellectual predecessors were smart enough to figure
that out 30 years ago! My only question is why it seems
so hard for us to emulate such a sensible practice in our
current world of research.

REFERENCES

CHALL, J. (1967). Learning to read: The great debate. New York:
McGraw-Hill.

CLARKE, L.K. (1988). Invented versus traditional spelling in first
graders' writing: Effects on learning to spell and read. Research in the
Teaching of English, 22, 281–309.

DYKSTRA, R. (1967). Continuation of the Coordinating Center for
first-grade reading instruction programs (Report of Project No. 6-
1651). Minneapolis, MN: University of Minnesota.

DYKSTRA, R. (1968a). The effectiveness of code- and meaning-
emphasis beginning reading programs. The Reading Teacher, 22,
17–23.

DYKSTRA, R. (1968b). Summary of the second-grade phase of the
Cooperative Research Program in primary reading instruction. Reading
Research Quarterly, 4, 49–70.

LYON, G.R., & CHHABRA, V. (1996). The current state of science
and the future of specific reading disability. Mental Retardation and
Developmental Disabilities Research Reviews, 2, 2–9.

PRESSLEY, M., RANKIN, J., & YOKAI, L. (1996). A survey of the in-
structional practices of primary teachers nominated as effective in pro-
moting literacy. Elementary School Journal, 96, 363–384.

SIPAY, E. (1968). Interpreting the USOE cooperative reading stud-
ies. The Reading Teacher, 22, 10–16.

SOUTHGATE, V. (1966). Approaching i.t.a. results with caution.
Reading Research Quarterly, 1, 35–56.