Contract No.: ED-01-0039/002
MPR Reference No.: 8911-700

MATHEMATICA
Policy Research, Inc.

# Assessing the Effectiveness of Education Interventions: Issues and Recommendations for the Title I Evaluation

***May 17, 2004***

***Steven Glazerman***
***David Myers***

Submitted to:

U.S. Department of Education
Office of Education Evaluation
555 New Jersey Avenue NW,
Room 308
Washington, DC 20208

Project Officer:
Audrey Pendleton

Submitted by:

Mathematica Policy Research, Inc.
600 Maryland Ave., SW, Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director:
David Myers

*PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING*

# ACKNOWLEDGMENTS

The authors thank Audrey Pendleton for her suggestions and encouragement throughout the design effort and the members of the Technical Working Group (TWG) assembled to provide guidance for this research design effort. Tom Trabasso and John Guthrie, who are also members of the TWG, and Michael Kamil provided valuable written suggestions. Several MPR staff contributed to this effort in "brainstorming sessions," including Allen Schirm and Mark Dynarski.

*PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING*

# TABLE OF CONTENTS


| Chapter | | Page |
|---|---|---|
| | **EXECUTIVE SUMMARY** | vii |
| **I** | **STUDYING INNOVATIVE PRACTICES UNDER TITLE I** | 1 |
| | A. BACKGROUND | 1 |
| | B. RESEARCH QUESTIONS AND GENERAL DESIGN CONSIDERATIONS | 2 |
| | C. PLAN OF THE REPORT | 3 |
| **II** | **SELECTING INTERVENTIONS AND SCHOOLS** | 5 |
| | A. SELECTING INTERVENTIONS | 5 |
| | 1. Specific Intervention Versus Intervention Types | 5 |
| | 2. Selecting Interventions | 6 |
| | 3. Feasibility | 12 |
| | B. SELECTING SCHOOLS | 13 |
| | 1. Defining Eligible Schools | 13 |
| | 2. Identifying Candidate Schools | 14 |

**Chapter** **Page**

**III EVALUATION DESIGN ISSUES** ..... 17

A. UNIT OF RANDOM ASSIGNMENT ..... 17

1. Randomly Assign Students ..... 17
2. Randomly Assign Classrooms ..... 18
3. Randomly Assign Schools ..... 18

B. STATISTICAL POWER AND SAMPLE SIZE REQUIREMENTS ..... 19

1. Methods and Assumptions ..... 20
2. Sample Sizes and MDEs ..... 21
3. Number of Districts ..... 21
4. Power of Subgroup Analysis ..... 23

C. DATA COLLECTION ..... 23

D. ANALYSIS ..... 26

REFERENCES ..... 29

APPENDIX A: TWG AND IRP MEMBERS ..... A-1

APPENDIX B: TEACHING OF COMPREHENSION IN CONTENT AREAS: DESIGN CRITERIA (BY THOMAS TRABASSO) ..... B-1

APPENDIX C: DESIRABLE CHARACTERISTICS OF COMPREHENSION INTERVENTIONS (BY MICHAEL KAMIL) ..... C-1

# T A B L E S


| Table | | Page |
|---|---|---|
| II.1 | EXAMPLES OF COMPREHENSION INTERVENTIONS | 7 |
| III.1 | ESTIMATED SAMPLE SIZE REQUIREMENTS AND MDE FOR FOUR TYPES OF HYPOTHESIS | 22 |

*PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING*

# F I G U R E S


| Figure | | Page |
|---|---|---|
| III.1 | MDE by Number of Students Per School | 24 |

*PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING*

# EXECUTIVE SUMMARY

Title I of the No Child Left Behind Act of 2002 (NCLB) is the largest funding vehicle of compensatory elementary and secondary education programs for disadvantaged children. This legislation calls upon educators to close the gap between low and high achievers by using instructional approaches shown to be effective by scientifically based methods. However, a limited knowledge base on the effectiveness of most instructional approaches has made it difficult for state and local educators to decide how to best use Title I funds to improve the educational outcomes of economically disadvantaged students. In October 2002, the Institute of Education Sciences (IES) contracted with Mathematica Policy Research, Inc. (MPR) to help identify issues pertinent to the evaluation of Title I and to propose feasible evaluation design strategies. The design effort took its lead largely from two sources: (1) the Title I Independent Review Panel (IRP), which was set up by Congress to provide ED with policy recommendations on Title I research, and (2) a more specialized technical working group (TWG) composed of education research and evaluation experts assembled for the design effort itself.

The MPR design team worked with the TWG, the IRP, other outside experts, and IES to develop a series of recommendations concerning (1) the topic of study for the evaluation, (2) strategies for selecting interventions and schools, and (3) the basic elements of a research design. Below we list the key recommendations, by area, that came out of this effort. The report itself provides the rationale for the recommendations and discusses design and measurement issues, including recommended sample sizes.

## EVALUATION FOCUS

- Reading comprehension should be the focus of the evaluation—particularly interventions designed to improve the reading comprehension of struggling students in grades three to five so those students can make progress in content areas such as social studies and science.

- The study should be designed so that reliable inferences can be made about selected groups of students, such as students with limited English proficiency and students with very low reading skills.

## A Process for Selecting Interventions and Schools

- A three-stage process should be used to select interventions. The first stage would involve soliciting proposals from the field. The second stage would involve winnowing down the proposals to a manageable number according to a set of initial criteria related primarily to whether it is feasible for the intervention developers to participate and to the ability of the interventions to incorporate social studies and science as content areas. In the third stage, a panel of experts would assess the finalists more critically based on factors related to their promise as effective interventions.

- Schools should be selected such that the evaluation sample is geographically diverse and representative of a range of schools with different concentrations of students with limited English proficiency. Furthermore, schools should generally have a high concentration of economically disadvantaged students.

## Key Research Design Parameters

- We recommend that *schools*, rather than classrooms or individual students, be randomly assigned to receive an intervention (treatment) or no intervention (control).

- Our power analysis suggests that, under reasonable assumptions, a sample of about 100 schools can achieve the evaluation objectives. These include the ability to test four interventions to determine whether they are effective and whether some are more effective than others.

- We recommend spreading the 100 schools across approximately eight districts.

- The in the NCLB implies that state assessments should be used to gauge student performance. We recommend that IES consider supplementing the analysis of state assessment results with a standardized test that measures students' achievement in reading, mathematics, science, and social studies.

# CHAPTER I

# STUDYING EDUCATION INTERVENTIONS UNDER TITLE I

## A. BACKGROUND

Title I of the No Child Left Behind Act of 2002 is the largest funding vehicle of compensatory elementary and secondary education programs for disadvantaged children. The legislation calls upon educators to close the gap between low and high achievers by using instructional approaches shown to be effective by scientifically based methods. However, a limited knowledge base on the effectiveness of most instructional approaches has made it difficult for state and local educators to decide how to best use Title I funds to improve the educational outcomes of economically disadvantaged students.

In October 2002, the Institute of Education Sciences (IES) contracted with Mathematica Policy Research, Inc. (MPR) to help identify issues pertinent to the evaluation of Title I and to propose feasible evaluation design strategies. The design effort took its lead largely from two sources: (1) the Title I Independent Review Panel (IRP) and (2) a more specialized technical working group (TWG). The Title I IRP was established by the Secretary of Education to to provide advice on methodological and other issues that may arise in connection with of the National Assessment of Title I. The TWG, which was formed by the design team and in consultation with IES, was made up of education research and evaluation experts assembled for the design effort itself. Members of the IRP and the TWG are listed in Appendix A. In addition to these two groups, MPR tapped the expertise of several consultants for background information on instructional strategies for enhancing skills in reading and mathematics. Consultants provided input via formal presentations to the IRP, conference calls with IES staff, and commissioned papers.

In its initial meetings, the IRP considered a wide range of topics that might be evaluated under the National Assessment of Title I, including comprehensive school reform, remedial reading, reading and math curricula, and professional development. To facilitate the discussion about potential topics, the IRP asked experts in reading and mathematics to provide an overview of the current state of research on the effectiveness of different reading and math curricula and of interventions for closing the gap between students with low and high reading and math achievement. Presentations on the subject were made by Drs. Jack Fletcher, Barbara Foorman, and Russell Gersten.

After considering these and other topics, the IRP selected reading as a key target for funds from the National Assessment of Title I. Following up on the IRP's recommendation, the TWG suggested that ED should narrow its focus to reading comprehension in content areas, such as science and social studies, for students in grades three to five. The decision to focus on reading comprehension, specific content areas, and students in grades three to five reflects the fact that (1) IES was already devoting considerable effort to understanding the effectiveness of reading programs for younger children and (2) even if these programs were effective, many disadvantaged children may still be struggling readers as they entered the higher elementary grades. The results of these discussion pointed the design team toward outlining a plan for and making recommendations about an evaluation of promising reading comprehension interventions for struggling readers in the middle elementary grades.

## B. Research Questions and General Design Considerations

A key question posed by the Title I IRP was: Do the practices supported by or potentially supported by Title I funds help students served under Title I meet or exceed state achievement standards? Conversations between MPR and the IRP, TWG, and IES suggest that this broad policy question might be better addressed if it were restated as a series of research questions that would frame an evaluation:

- What are the impacts of the selected interventions or instructional approaches on reading comprehension for Title I-eligible students in grades three to five? What are the impacts relative to prevailing practices? What are the impacts relative to other interventions?

- What are the impacts of the selected interventions or instructional approaches on reading comprehension, achievement in content areas that require reading comprehension, and achievement in other areas?

- What are the impacts of the selected interventions or instructional approaches for different types of students, such as students with limited English proficiency and students with learning challenges?

Addressing these questions adequately requires that we have a common understanding of the meaning of the term "impact" in the context of an evaluation. "Impact" implies a causal relationship between a treatment and an outcome. The most convincing way to establish causality is through a controlled experiment in which random assignment is used to construct treatment and control groups to determine whether and the extent to which reading comprehension interventions have an effect, or an impact, on achievement. Because random assignment makes the treatment and control groups—in this case, of students, classrooms, or schools—statistically equivalent with the exception of the opportunity to participate in an intervention, a controlled experiment allows us to conclude that differences in outcomes between the treatment and control groups (i.e., the impacts of the treatment) were caused by the treatment.

The TWG suggested that an experimental design should be feasible to address the questions posed in the Title I IRP and as revised by the TWG, meaning that one should be able to find real-world situations where students, classrooms, or schools that can be randomly assigned to intervention and nonintervention conditions.

Random assignment is a good tool for achieving "internal validity," meaning that the difference in outcomes between treatment and control groups can be properly attributed to the intervention. Although random assignment minimizes the threats to internal validity more so than do other research designs, it does not eliminate all threats, such as imitation of treatments, compensatory equalization, compensatory rivalry, and demoralization in groups receiving less desirable treatments (see, for example, Cook and Campbell 1979). On the whole, however, experimental designs produce more reliable and valid results than do quasi-experimental or nonexperimental approaches, which are vulnerable to a greater variety of threats (for a full accounting, see Cook and Campbell 1979).

In addition to considering whether an experimental or non-experimental design should be used to assess the impacts of reading comprehension interventions, one must consider whether it should be an *efficacy study* or an *effectiveness study*. The former is used to assess an intervention under laboratory-like, or highly controlled, conditions in order to determine whether the intervention might possibly work under the best of circumstances. The latter is used to assess an intervention under real-world conditions that are, by definition, not as well controlled in order to determine whether the intervention is likely to work under typical circumstances. For example, professional development programs for teachers who are being introduced to an intervention may vary from one school setting to another because of variation in trainer competence. As another example, some teachers may adapt the intervention curriculum or instructional methods to their perception of what is appropriate for their students. Failure to control these conditions could be said by some to dilute the intervention's impacts. On the other hand, those impacts might be closer to what one would expect when the intervention is implemented on a larger scale—or, in other words, a more accurate indication of the intervention's effectiveness.

Research on interventions often progresses first from efficacy studies to prove a concept to effectiveness studies that are used to justify more widespread adoption of the intervention. Despite what may be a limited number of reading comprehension interventions with substantial evidence about their efficacy, the TWG recommended that the evaluation of reading comprehension interventions for Title I should be an effectiveness study.

## C. PLAN OF THE REPORT

Chapter II recommends procedures for selecting interventions and schools for the study. Chapter III discusses random assignment designs, sample size requirements, data collection plans (including student achievement measurement), and data analysis.

*PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING*

# CHAPTER II

# SELECTING INTERVENTIONS AND SCHOOLS

Two major challenges for the study will be to select the reading comprehension interventions and to secure the participation of schools where the interventions will be implemented. These tasks are challenging because both the interventions and the schools need to satisfy a number of criteria that make them suitable for an effectiveness study. Based on input from the IRP, the TWG, and IES, this chapter provides recommendations on the selection criteria and how they may be used to arrive at a final set of interventions and schools for the study.

## A. SELECTING INTERVENTIONS

The legislation calling for the U.S. Department of Education (ED) to evaluate effective practices under Title I does not specify what those practices should be. As discussed in Chapter I, ED has solicited input from a variety of sources to narrow down the problem to specific populations and pedagogical challenges. There still remain several issues in identifying the interventions that target those challenges and populations. These issues include deciding whether to evaluate specific interventions or broad classes of interventions and deciding which interventions among the many candidates should be included.

### 1. Specific Intervention Versus Intervention Types

An issue raised in the TWG meetings and in discussions with IES was whether the study should focus on examining the effect of *named, packaged* interventions, such as those marketed by major publishers and those developed by university-based researchers, or on the effect of several *types* of interventions, each having a distinct approach or a distinct combination of instructional components.

We recommend that IES select the most promising named, packaged interventions. The rationale for this recommendation is that reading experts on our TWG did not agree on a single typology that could be used objectively and without controversy to categorize the diverse set of approaches that might be offered as candidates for further study. Classification schemes for comprehension interventions can be found in the report of the

National Reading Panel (National Reading Panel 2000) and in other places, but for every scheme, there are often interventions that straddle two or more categories.

Moreover, an intervention selected on the basis of type may, when implemented in a real classroom setting, look different from what was expected according to how it was classified. Grouping and then selecting interventions by type is more appropriate in settings such as the ongoing Education Technology study, where programs can be grouped by easily distinguishable content areas (math versus reading) and grade levels (grade one versus grade five). In other settings, such as the Preschool Curriculum Evaluation Research (PCER) study and in the proposed study, where there is a single age-cohort and a specific set of cognitive skills being targeted, it becomes difficult to develop a typology of interventions in a way that is objective and widely acceptable. It will be easier to classify specific interventions after they have been selected and observed in the field. Preliminary judgments about the types of interventions and key components can be used during the selection process to ensure that a diverse set of interventions is selected.

The set of "named interventions" does not have to include only curricula that are published commercially. Some interventions, such as Reciprocal Teaching, are widely disseminated and practiced by individuals who may be affiliated with different organizations. Published guides may be detailed enough so that practitioners will follow the approach consistently.

## 2. Selecting Interventions

We recommend a three-stage process for selecting interventions. The first stage would involve soliciting proposals from the field. The second stage would involve winnowing down the proposals to a manageable number according to a set of initial criteria related primarily to whether it is feasible for the intervention developers to participate and to the ability of the interventions to incorporate social studies and science as content areas. In the third stage, a panel of experts would assess the finalists more critically based on factors related to their promise as effective interventions.

### a. Invite Bids from Intervention Developers and Publishers

To ensure that all eligible interventions are considered for the evaluation, one would ideally start with a list of all possible interventions; however, such a list does not exist. Therefore, through literature searches, consultation with members of the TWG, and over a dozen interviews with curriculum developers and publishers, MPR developed an initial working list of interventions that may satisfy most of the selection criteria (see Table II.1).

**Table II.1. Examples of Comprehension Interventions**

| Intervention | Developer or Publisher |
|---|---|
| CORI (Concept Oriented Reading Instruction) | John Guthrie (U. Maryland) |
| QtA (Questioning the Author) | Isabel Beck (University of Pittsburgh) |
| Reciprocal Teaching | Ann Marie Palincsar (University of Michigan) |
| SAIL (Students Achieving Independent Learning) | Michael Pressley (SUNY Albany) |
| PALS (Peer Assisted Learning Strategies) | Doug Fuchs and Lynn Fuchs (Vanderbilt) |
| Theme Scheme/ Expository Text Comprehension | Joanna Williams (Teachers' College, Columbia) |
| Four Blocks/Guided Literacy | Pat and Jim Cunningham and Dorothy Hall (Wake Forest) |
| Language and Literacy Framework | Irene Fountas (Lesley College) & Gay Su Pinnell (Ohio State) |
| Collaborative Reasoning From Text | Richard Anderson (U. Illinois-Champaign) |
| Reading Mastery Plus, Corrective Reading, Open Court | SRA/McGraw-Hill |
| Soar to Success | Houghton-Mifflin |
| Visualizing and Verbalizing | Lindamood-Bell |
| Core Knowledge | Core Knowledge Foundation |
| Reading Wings | Success for All Foundation |
| Universal Design for Learning | Center for Applied Special Technology (CAST) |
| Comprehension Upgrade | Learning Upgrade |

The selection criteria include the following:[1]

**Targeted.** Based on recommendations from the IRP, the TWG, and discussions with IES, we believe the intervention should address the comprehension skills of students in at least one of the middle elementary grades (interpreted as three through five), be appropriate for English language learners (ELLs), and be able to encompass such content areas as science or social studies with minimal development work. This combination of age group/grade, skills set, and content focus was strongly endorsed by the TWG. Applicants who nominate interventions should be asked to document the following, with specific references to curriculum materials or teacher training materials where applicable:

[1] Many details in the list of criteria, including subcriteria, draw heavily on the two papers commissioned for this effort (Trabasso 2003; Kamil 2003). The papers are reproduced here in Appendixes B and C.

To show that the intervention targets the right population,

- Appropriateness for third-, fourth-, or fifth-grade students, with examples of prior implementation for at least one of these age groups (the intervention is not required to target all these grades)
- Types of background knowledge required of and taught to students through the intervention
- The nature of any accommodations for limited English proficient students

To show that the intervention targets the right skills,

- The comprehension strategy or strategies and how it/they are taught (e.g., implicit, explicit) and vocabulary instruction (e.g., implicit versus explicit)
- The role of oral language and word-level skills such as phonemic awareness and decoding in terms of both the initial skills expected of students and the program components intended to improve these skills
- Any program components that target student engagement and motivation

To show that the intervention has materials and components that support targeted skill development,

- A description of the expository texts (basal readers, trade books, etc.) or a description how the texts are selected and used, including a description of any multi-media methods
- Overall description of the types of tasks (such as writing), time on task, and amount of practice required, as applicable
- Any intervention-specific assessments (e.g., screening tools used to assess children's entry-level vocabulary)
- Other features that may be unique to the proposed intervention

For example, an intervention that is likely to meet the "targeted" criterion is Concept-Oriented Reading Instruction (CORI), which has been used to teach third- and fourth-grade students by using an engagement model (promotes student engagement with subject matter) to teach lesson units on the solar system or simple machines. In addition to CORI, other interventions may be able to satisfy the "targeted" criterion including Visualizing/Verbalizing, Guided Reading (a component of the Four Blocks intervention), and Open Court (which has been adapted in California schools to work with science and social studies).

Other interventions, such as Core Knowledge, Theme Scheme, and Reading Wings could be adapted to meet the criterion. Core Knowledge has content-rich texts and a teacher-directed focus but was not designed with a goal of explicit instruction of reading skills or strategies. Theme Scheme uses fictional narratives but can be applied to expository texts with some adaptation. Reading Wings includes materials on reading expository text and accompanying social studies and science trade books, but they are not the focus of the intervention. According to the developers, some development work would be required. The content focus for an adapted Reading Wings could come from WorldLab, a related intervention from the same developer, which includes an integrated science and social studies program.

**Promising.** The intervention should have some evidence of efficacy. If one thinks of the study of effective practices under Title I as an analog to a Phase III clinical trial in the life of a new drug application for approval by the U.S. Food and Drug Administration, the experimental interventions of interest here would have to show evidence at the level of Phase I or Phase II experiments that they do no harm and have some promise for boosting both reading comprehension and subject area knowledge. Because the evidence used to justify an intervention's efficacy might be nonexperimental or dependent on inferences from small samples, it would be important to have experts in statistical inference and research design carefully assess the quality and relevance of the evidence.[2] To facilitate this process, applicants should be asked to provide:

- A list of experimental studies and their findings
- A list of quasi-experimental studies and their findings
- A list of other types of evidence

If the studies pertain to interventions that have been modified since they were tested or implemented with an idiosyncratic sample of students or students who are at a different developmental stage than the target population for the proposed study, then applicants should include a brief statement with each piece of evidence to explain how the research is relevant to and supports the proposed intervention. The panel of experts can evaluate claims that, for example, an intervention that is proven effective for first grade students would be presumed promising for third graders, or that an intervention that combines three strategies that were proven effective on their own would also be effective in combination.

There may exist contrary evidence that was not volunteered by the intervention developers that IES could consider in its decision about whether an intervention is promising. We recommend that the experts chosen to review the interventions be selected from among those who have a broad knowledge of the research literature. In that way, IES can be made aware of any interventions that have been "proven" in an isolated study to be effective, but a preponderance of evidence exists that refutes the developer's claims.

[2] Instruments developed by IES' *What Works Clearinghouse* may be used to assess the quality of evidence.

Our preliminary investigation found that there is a wide variation in the interventions' ability to back up their approach with research. Reciprocal Teaching is one of the few that has been the subject of a large volume of research. CORI is one of the few that has both quasi-experimental and experimental evidence of effectiveness. Other interventions—such as Comprehension Upgrade, Reading Wings, and Guided Reading—may not have an extensive research base on their own but were developed as extensions to other interventions—Reading Upgrade, Roots and Wings, and Four Blocks, respectively—that have been researched more rigorously. We recommend that IES take this type of indirect evidence into account but give it less weight in determining promise, and include the intervention only if there is a logical connection between research on a related intervention and the intervention being proposed.

**Replicable**. Interventions included in the evaluation should be replicable, meaning they should be well defined and able to be implemented consistently with reasonable fidelity to the program model. Intervention developers should be able to:

- Document how fidelity is measured
- Provide fidelity data for past implementations if applicable
- Provide fidelity measurement instruments on request

**Scalable.** A scalable intervention can be implemented in many schools without altering its fundamental nature. An intervention that might *not* be scalable would require a single person, such as the intervention developer, to provide all training in order for the intervention to be properly implemented. Interventions that are supported by networks of trainers or that have a systematic process of training the trainers, such as Four Blocks (Guided Reading) or Reading Wings, are more likely to be scalable.

To address the issue of scalability, developers can be asked to report on:

- How many classrooms and schools have adopted the intervention to date
- How the intervention is delivered to the schools, e.g. on-site professional development, on-site coaching, off-site coaching via email and telephone, developer-prepared materials, technology applications

### b. Develop a "Short List"

After inviting developers to nominate interventions for the study, the next step would be to winnow down the applications to a set of finalists. We recommend that IES, in taking this step, make several "passes" with assistance from an evaluation contractor. The first pass should be based on the most easily and objectively evaluated factors. For example, it could eliminate interventions that are not appropriate for the target grade levels, for low-income students, or for students with limited English proficiency. Replicability and

scalability criteria can then be applied in subsequent passes through the set of candidates, as can the reading skills and subject matter criteria.

One last pass before selecting finalists would address the costs of implementation. Our informal survey revealed that typical interventions could cost roughly $20,000 to $40,000 per school (three classrooms) including training and materials. Developers/publishers of some interventions, such as Questioning the Author, Reading Mastery Plus, and Comprehension Upgrade, provided lower cost estimates, ranging from about $5,000 to $12,000 per school. However, these estimates may not include all the fixed costs. Some intervention developers, such as the Success for All Foundation, estimated that costs would be considerably higher for a startup year. Other fixed costs might include a pilot phase for developing a subject matter unit or for purchasing equipment for multimedia applications. At this stage of the selection process, IES would essentially be seeking to eliminate interventions whose costs are prohibitive, both because of the burden on the study and, even if it could be funded for the study, the difficulty Title I schools with limited resources would face in funding the intervention on an ongoing basis.

### c. Select Finalists

We recommend convening a panel of experts in reading comprehension and program evaluation to guide the final selection of interventions. The panel would make judgments on the more complex criteria, such as whether the theory and logic of the interventions are coherent and whether the intervention is sufficiently promising based on the research evidence. The panel could also offer recommendations to IES about the broad types of approaches to ensure that multiple options are tested. If the interventions are too similar to one another, they should be grouped together, considered a single type of intervention.

An important issue in selecting the finalists is the reliability of the effectiveness evidence. Few interventions have been evaluated using rigorous methods except on a small scale. A cursory examination of the evidence put forward by developers and publishers we interviewed showed only a few studies based on experimental or sound, appropriate quasi-experimental methods.

To determine whether a particular piece of evidence is reliable, the design team will develop a rubric that the expert panel can use to rate the degree to which a study provides causal evidence of the effect of an intervention. A study that is rated as providing strong causal evidence will have addressed all the issues associated with comparing outcomes for individuals in the intervention (treatment) group to outcomes for those in the control (or comparison) group. The rubric will describe the aspects that should be examined, such as the type of evaluation design, quality of the data, and whether the evidence is internally and externally valid. It will also provide guidance on how much weight each aspect of study quality will receive, allowing one to rate the evidence base according to a score.

Two rubrics now being developed may be useful in the study. One—developed as part of ED's What Works Clearinghouse—is the Study Design and Implementation Assessment Device (DIAD). The other, now being developed by a committee of Division 16 of the

American Psychological Association, is referred to here as the APA coding criteria. The strength of both lies in their focus on research about the causal effects of educational interventions.[3] (Several articles on the APA coding criteria appear in the December 2002 issues of the *School Psychology Quarterly*.

### 3. Feasibility

The feasibility of the proposed study depends on whether there are interventions that can meet the criteria listed above and on whether developers of such interventions will be willing to participate in the study. Our conclusion from interviews with developers and publishers is that the intervention selection plan—which calls for inviting intervention developers to apply for grants to implement research-based interventions as part of a rigorous, random assignment impact study of reading comprehension in the content areas—is feasible. Many respondents suggested that they would be able to satisfy most, but not all, of the criteria. For example, some were concerned about being able to scale up to the number of schools required for the study. Another developer reported that the research backing up the intervention was done with a different age group or a different version of the curriculum. For others, the concern had more to do with integrating content areas into the intervention. Therefore, a reasonable approach to selection would be to retain the criteria as evaluative factors, not requirements, so as to invite more applications and leave more selection options open.

A feasibility issue specific to studying multiple reading comprehension interventions, especially those promoting reading in science and social studies, is whether the reading techniques can be adapted successfully to content area lessons as opposed to a pure language arts lessons. This problem amounts to selecting texts for use by the classrooms under study. Reading comprehension is often taught with fictional narratives, but the type of learning recommended by the TWG members, who emphasized background knowledge and subject mastery, would require expository texts. Not all reading comprehension intervention developers will have addressed this issue.

It may be possible for IES to economize by supporting one separate effort to select trade books (expository texts) that can be used for all interventions. This would achieve two purposes simultaneously: it would save on development costs, so multiple authors would not duplicate the same task, and it would allow IES to hold constant an important variable in the impact study. We caution only that a centralized text-selection process should be proposed as an *option* to be exercised only if it would be appropriate given the interventions that are ultimately selected. Some programs, such as CORI, may rely on teachers' flexibility in selecting texts as a key component of the intervention itself.

[3] The DIAD is published on the Internet by the What Works Clearinghouse at www.w-w-c.org/DIAD_Final.doc, most recently accessed on April 19, 2004.

## B. SELECTING SCHOOLS

### 1. Defining Eligible Schools

To be eligible for the study, schools must both provide a real-world setting that is conducive to supporting a well-implemented intervention and support policy goals such as improving student achievement as part of receiving Title I funds. The schools must also support the study's legislative mandate to use rigorous methods, including control groups and random assignment, to the extent feasible, to estimate effects. Suggested criteria for selecting schools and school districts are listed below. The first list applies to each school under consideration. The second list applies to the final set of schools selected to participate.

- ***High Poverty.*** The goal of the Title I program is to help the nation's most disadvantaged children meet state standards for academic achievement. Therefore, the evaluation would be most policy-relevant if we identify interventions that are effective for the most disadvantaged students. Setting a threshold, such as schools having at least 40 percent of the students in the target grade eligible for free or a reduced-price lunch would achieve this goal. An alternative, simple way to identify high-poverty schools is to select those that operate a school-wide program under Title I.[4]

- ***Not Meeting Adequate Yearly Progress (AYP) Targets.*** A schools' failure to meet its AYP goals, while not explicitly listed as a criterion in the legislation authorizing the proposed Title I research, could be an incentive for the school to adopt policy change. Participation in the reading comprehension study could be offered as a way to satisfy the corrective action requirement under the No Child Left Behind law.

- ***Willing to Adopt Interventions.*** It is generally understood that successful school reform requires the buy-in of school staff. As part of the study, IES can provide tangible participation incentives to school district officials, principals, or teachers. Principals and teachers in the target grades in particular will have to be directly convinced of the value of the study.

Once individual schools' eligibility is determined, the following criteria should be applied to the full set of schools, rather than to any one school in particular.

- ***Geographic Diversity.*** While it may not be possible or desirable to draw a nationally representative sample of schools, it is still important to achieve some degree of external validity by selecting schools that are not too geographically concentrated in one area. A purposive sample of districts that covers states with

[4] Title I funds are targeted to low-income students. School-wide programs are those that provide Title I funds to an entire school if more than 40 percent of the students would be eligible individually.

a variety of policies and regional characteristics would make it likely that the interventions are being tested in a range of settings and the findings thus applicable to a diverse set of schools that would consider adopting a successful intervention.

- ***Language Proficiency and Ethnic Minority Representation.*** While geographic diversity and a low-income student population may produce a representative study sample, it is important to ensure that ethnic minorities, and limited English proficient students (LEP), are also represented. The Title I IRP singled out LEP as a group that should be represented in the study. This implies that a sizable fraction of schools should have some representation of LEP in order to support subgroup analysis. If possible, it would be desirable to have some school districts with a small percentage of LEP and some with a high percentage of LEP so that we could observe whether interventions are effective for LEP in each type of setting.

- ***Clustering of Schools in Districts.*** As discussed in the next chapter, we recommend that the study design use random assignment of schools and that the schools be concentrated in a small number of districts. It is critical for the success of this random assignment design to select school districts such that multiple schools with similar characteristics can participate. Ideally, this would be achieved by finding multiple elementary schools within the same district, although a "virtual district" could be formed by using schools from different districts if the policies, populations, and curricula were similar.

One factor that MPR considered as a possible criterion was whether the school was implementing Reading First, the federal government's program to promote core reading skills in the early elementary grades. We believe that it would be useful to observe the effectiveness of comprehension interventions in both types of settings, those with and without Reading First programs. For those schools that have Reading First programs in place, the proposed comprehension interventions may be seen as a follow-on, to continue with innovative practices that bring reading skills developed in early elementary grades to enhance those skills in the middle elementary grades.

## 2. Identifying Candidate Schools

Once the eligibility criteria are defined, two important steps would be to: (1) find a sample of schools that meet the criteria and (2) induce them to participate in a random assignment study.

There are several approaches to identifying a sample of schools. With the assistance of an evaluation contractor, IES could search a public use dataset such as the Common Core of Data and other federal and state sources to identify the universe of school districts with the requisite number of elementary schools that meet the poverty criterion but not AYP targets. Only schools that have at least grades three through five should be considered so that evaluators do not have to track students as they fan out to different schools during the

followup period if the study is to last more than one year. Once this universe is identified, a geographically stratified sample of states and then districts can be drawn, and the sampled districts can be invited to participate.

Another way to identify schools would be to poll membership organizations that have close contact with the universe of school districts that are likely to meet the study criteria. For example, the Council of Chief State School Officers could be helpful in identifying and recruiting school districts for the study.

Making participation voluntary at the district level defines a smaller universe of schools to which study findings can be generalized but does not compromise internal validity. Furthermore, defining the universe in this way is acceptable given that it is impractical to conduct the study in a very large number of districts to begin with. Understanding the effects of interventions in districts that volunteer and hence are more eager to adopt new interventions is arguably more policy relevant than knowing about impacts for a Title I school in a randomly selected district.

Securing participation can be as challenging and perhaps more challenging than identifying schools. Data collection requirements and compliance with random assignment when each school has a chance of being assigned to receive no intervention are likely to be seen as burdens and therefore present barriers to schools' willingness to take part. The extent of the required recruiting effort depends on the size of the incentives. If the study schools are concentrated in a relatively small number of districts, IES would be able to offer larger amounts of intervention support to each district, making participation in the study very attractive.

Some additional funds should probably be included for control schools to implement the intervention of their choice either in a nontargeted grade level or in the year following the implementation year, as is sometimes done in well-designed experiments (Myers and Dynarski 2003). Such funding (and flexibility) for control schools help to gain their cooperation and maintain fairness. In addition, IES could offer to fund the salary of a local research coordinator, typically a member of the district's administrative staff, who could serve as a liaison to the study and facilitate both data collection and monitoring of compliance with the study protocol.

*PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING*

# CHAPTER III

# EVALUATION DESIGN ISSUES

The success of the proposed study will depend in part on how well the evaluation design reflects the study goals. The most important design decisions are related to the unit of random assignment (students, classrooms, or schools), the number of units at each level to set as sample size targets, and the data collection and analysis plans.

## A. UNIT OF RANDOM ASSIGNMENT

Random assignment is a flexible design tool that, in principle, can be used to assign students, classrooms, or entire schools to treatment and control groups. We recommend that *schools* be the unit of random assignment. This section describes the strengths and weaknesses of various assignment options and the rationale for choosing schools as the unit of assignment.

### 1. Randomly Assign Students

The most statistically efficient research design would randomly assign individual *students* to treatment conditions.[1] Unfortunately, this approach is appropriate only if the intervention could be administered to some students and not to others within the same class. This is possible only for some reading interventions—for example, pullout tutoring programs like Reading Recovery or individualized instruction programs like Learning Upgrade, in which children receive self-guided instruction at computer terminals. Reading experts on the TWG suggested that the interventions with the best hope of making a real difference in Title I students' reading comprehension skills are those that change the instruction in the entire classroom. This rules out pullout programs and suggests that intervention at the teacher/classroom level is preferred.

[1] "Statistically efficient" is used here in a general sense that it is easier to generate large samples of independent observations with smaller units like students than it is with larger units like classrooms or schools. Larger samples of independent replications of the treatment increase the statistical power of the study's hypothesis tests.

### 2. Randomly Assign Classrooms

In the next most efficient design, *classrooms* would be randomly assigned to one of the reading comprehension interventions or to no intervention. There are typically only a few classrooms per grade in any given school, but this unit of assignment would allow researchers to observe sample members assigned to different treatments in the same school. Varying treatment assignment at the classroom level is a powerful approach because it allows us to compare classrooms in the same school and thereby rule out the school climate and other school-level factors as an influence on student outcomes.

A disadvantage of randomly assigning classrooms within schools is known as "spillover" or "contamination" effects. This phenomenon typically happens when teachers are aware that one or more of their colleagues is delivering instruction differently or receiving some special intervention, and this awareness influences their behavior. These behavioral changes induced by the experiment were alluded to in Chapter 1 (imitation of treatments, compensatory rivalry, and demoralization). Spillover can also happen when students in intervention and nonintervention classrooms interact, possibly closing the gap between their differences and thus their outcomes. In either case, the impact estimates would usually be biased toward zero.

Cost efficiency and fairness argue against random assignment of classrooms. Many developers of reading comprehension interventions that we consulted for this report suggested that implementation costs are driven more by the number of schools than the number of classrooms in a study. For example, a single teacher training session can be offered to all teachers in the target grade for a given school, and return visits for periodic teacher coaching during the school year might be charged on a per-school basis. This means that intervening in a large number of classrooms is more expensive if the classrooms are spread out among schools than if they are clustered within schools, as discussed below.

Fairness also argues against assigning classrooms. For instance, some principals may be reluctant to have their school participate in a study in which some teachers are provided with special training or materials and others are not. Even if principals allowed the differential treatment within a school, there might be pressures to allow some practices to spill over into nonintervention classrooms, thus biasing impact estimates. There might also be pressures to allow students perceived to "deserve" one treatment over another to transfer (cross over) to the "classroom of interest," also biasing the impact estimates.

### 3. Randomly Assign Schools

We recommend randomly assigning *schools* to treatment groups. This approach will require a larger sample of schools than the other approaches in order for the evaluator to disentangle the treatment effects from school-level characteristics. The larger sample of schools can be costly because of the fixed costs associated with securing schools' cooperation. However, the added resources would be a good investment, as the random assignment at the school level will eliminate many of the threats both to the study's feasibility and its internal validity.

To increase the precision of the impact estimates and to reduce the number of schools needed for the evaluation, we recommend a randomized block design, which is analogous to stratification techniques used to make statistical sampling more efficient. One blocking technique is to first identify schools that can be paired or grouped according to similarities in the characteristics that are considered crucial to outcomes and then conduct random assignment within pairs or groups. Blocking factors could include the experience levels of the teachers or the percentage of students who are eligible for a free or reduced-price lunch.

Another blocking factor that we believe is critical is the district. That is, the evaluation team should conduct random assignment of pairs or groups of schools within districts in order to hold district policies such as teacher hiring, compensation, and professional development constant. Most important, conducting random assignment within a school district would hold constant the curriculum and standard texts used in the classroom. This approach follows Trabasso's (2003) recommendation that curriculum and texts are critical factors to control.

## B. Statistical Power and Sample Size Requirements

To ensure that the evaluation will both address the research questions with sufficient statistical precision and stay within budget, IES and its evaluation contractor need to develop sample size targets for students, classrooms, and schools. Sample size requirements can be estimated by applying conventional formulas to a set of assumptions about the structure of the particular experiment and incorporating some additional assumptions about the variability of the outcomes of interest. Following the recommendation of IES, we have based the power analysis on a study with four interventions. We have further assumed that random assignment will be done at the school level, blocked (stratified) by district. The research questions include the following:

1. What is the impact of each intervention relative to the status quo (control)?

2. Do some interventions have larger impacts than others?

3. Do the impacts differ for subgroups of students?

Question 1 can be addressed by comparing the average outcomes for schools assigned to each intervention group with the outcomes for the schools assigned to the control group. Question 2 can be addressed by comparing outcomes for each intervention group to one another. Question 3 can be addressed by repeating the analysis for questions 1 and 2, but with subsets of the data defined in terms of student or school background characteristics.[2]

[2] More detail on analysis strategies is presented in Section D below.

## 1. Methods and Assumptions

The presence of multiple interventions and the desire to make multiple comparisons should be accounted for in the power analysis. The conventional approach sets a target of 80 percent power and 10 percent significance level (probability of falsely concluding there is significant difference). As the number of comparisons goes up, however, so does the probability of finding a "statistically significant" impact purely by chance. To avoid this problem, we recommend using multiple comparison procedures called the Dunnett's t-test and the Duncan multiple range test. Both provide an appropriate appraisal of the significance level, but they call for slightly larger samples than would be needed if only one intervention was being studied.

Dunnett's t-test, which is used to test each of the interventions against the control (question 1) is similar to the conventional t-test but uses a larger critical value. Duncan's procedure, used to test the interventions against each other (question 2) allows us to achieve lower minimum detectable effects (MDEs) than the usual approaches (Bonferroni adjustments, for example) because it specifies an order in which hypotheses are tested. It calls for a sequence of tests to be conducted as follows. The four treatment means are "lined up" from smallest to largest and labeled A through D. The test is performed for the comparison of A to D using a critical value found in a special lookup table based on the number of interventions being compared. If that test is rejected, meaning the difference is statistically significant, then tests are performed for A vs. C and B vs. D. If those are rejected, then each of the "adjacent" pairs is tested (A vs. B, B vs. C, and C vs. D). The critical values for each test, which depend on the number of rank positions between the two treatments being compared, can be found in a Duncan's multiple range test lookup table.

Two additional assumptions underlying the power analysis should be explained. We assume 90 percent of test score variance is explained within schools, with the remaining 10 percent being between-school variance. This 90/10 ratio is consistent with empirical evidence from data on elementary school achievement collected for the national evaluation of Teach for America. The other assumption is that 50 percent of the variance in outcomes (e.g., test scores) at the end of the observation will be explained by covariates such as baseline achievement level and student and family background. Again, this assumption is reasonable given the empirical evidence available from existing studies.

Another consideration for the power analysis is the ratio of schools assigned to each of the treatment groups. As a general rule, a balanced design, in which the same number of schools is assigned to each treatment group, will be the most statistically efficient. This rule does not apply when multiple treatments are compared with one control group, however. In such cases, we can do better by assigning $N_c$ units to the control group and $N_t$ units to each treatment group such that the ratio $N_c/N_t$ is proportional to the square root of the number of treatment groups (Box, Hunter, and Hunter 1978)—in this case, five.

The decision of whether to use the same number of treatment and control schools, a ratio optimized for multiple treatment-control comparisons only, or a ratio optimized for pairwise treatment comparisons only depends on which research question IES wants to emphasize the most. For the power analysis, we assumed that answering Question 1

definitively is the first priority, with the opportunity to address Question 2 being a lower priority. If the primary goal were to answer Question 2 with a high level of precision, the size of the overall sample would need to be dramatically larger and the number of control schools would be proportionally smaller. To approximate the optimal ratio for multiple treatment-control comparisons, we considered sample size configurations with a ratio of 9 control schools to (4 schools x 4 interventions =) 16 treatment schools ($9/4 \approx \sqrt{5}$).

## 2. Sample Sizes and MDEs

Using this ratio, our power analysis suggests that a sample of about 100 schools will achieve the evaluation objectives (see Table III.1). The sample of 100 schools consist of 16 schools implementing all four interventions plus 34 control schools, with two classrooms in each school and 30 students per school after accounting for attrition and nonresponse, resulting in an analysis sample of about 3,000 students.[3] With this design, the study can detect an effect size (impact) of 0.25 for comparisons of an intervention with a control group. This threshold for policy relevance of one-quarter of a standard deviation, although somewhat arbitrary, represents a reasonable floor for considering an intervention to be "effective." The National Reading Panel (2000) reviewed rigorous studies of comprehension interventions and found effect sizes for impacts on student achievement that ranged from 0.24 to 1.70. The median effect size of the six studies of Reciprocal Teaching reviewed by the NRP was 0.34 when looking at standardized tests. For the seven studies of comprehension interventions, the median effect size was 0.35. Larger effect sizes were found with assessments that tested comprehension directly.

For the comparisons between treatments, the MDEs are somewhat larger. The MDE between the most and least effective interventions is estimated to be 0.50, and the MDE for the interventions that are two positions apart (the intervention with the highest scores versus the third-highest, or the second-highest versus the lowest) is 0.44. If both of these tests show significant differences, then the study will be able to detect differences of 0.34 between adjacent pairs.

## 3. Number of Districts

Still open is the question of whether the 100 schools should be spread across a large number of districts or concentrated in a smaller number of districts, with more schools per district. A large number of districts provides greater external validity but introduces a source of variation that is hard to control. We recommend limiting the number of districts as much as possible because achieving a sample that is representative of all districts would be

[3] All sample sizes refer to the number of units that will be used in the analysis. If some schools are expected to drop out of the study or fail to comply with their treatment assignment, then a higher initial sample size will be required. Similarly, if student attrition is expected, then the initial sample targets should be increased to offset the expected losses. For example, if 10 percent attrition is expected, then the target should be increased from 30 to 33 students per school.

**Table III.1. Estimated Sample Size Requirements and MDE[a] for Four Types of Hypothesis Tests**

| Number of Schools[b] | | | MDE by Type of Hypothesis Test[c] | | | |
|---|---|---|---|---|---|---|
| Treatment[d] | Control | Total | A vs. Control | A vs. D | A vs. C[e] | A vs. B[e] |
| 32 | 18 | 50 | 0.34 | 0.73 | 0.64 | 0.49 |
| 48 | 27 | 75 | 0.28 | 0.59 | 0.52 | 0.40 |
| **64** | **36** | **100** | **0.24** | **0.50** | **0.44** | **0.34** |
| 96 | 54 | 150 | 0.19 | 0.41 | 0.36 | 0.28 |
| 128 | 72 | 200 | 0.17 | 0.35 | 0.32 | 0.25 |
| 192 | 108 | 300 | 0.14 | 0.29 | 0.25 | 0.20 |
| 256 | 144 | 400 | 0.12 | 0.25 | 0.22 | 0.17 |
| 320 | 180 | 500 | 0.11 | 0.22 | 0.20 | 0.15 |

Assumptions

50% of variance in test scores is explained by covariates (including pretest)
90% of (residual) variance in test scores is within schools, 10% between schools
Hypothesis tests conducted with 80% power, α=0.10, one-sided;
30 students per school

Notes

[a]MDE = minimum detectable effect. All MDEs are shown in terms of effect size, which is equivalent to the percentage of a standard deviation in the outcome, divided by 100.

[b]The treatment schools are divided equally among 4 interventions. We assumed a ratio of 9 control schools to 4 treatment schools for any given treatment-control comparison. This approximates the optimal ratio for a design with four treatment groups and one control.

[c]The test of A vs. Control is equivalent to B vs. Control, C vs. Control, and D vs. Control and is based on Dunnett's procedure. Tests of treatments with each other assumes the interventions have been ranked by their mean outcome score and labeled alphabetically, so A vs. D is the test of difference between the highest and lowest mean. The test of A vs. C is equivalent to B vs. D and tests the difference between the interventions that are ranked two positions apart. The test of A vs. B is equivalent to any pairwise comparison of adjacently ranked interventions (B vs. C or C vs. D). These tests are all based on Duncan's Multiple Range procedure.

[d]The sample size listed is the number of treatment schools per intervention.

[e]Following Duncan's procedure, A vs. C type tests would only be performed if (A = D) is rejected, and A vs. B type tests would only be performed if (A = C) is rejected.

prohibitively expensive, so a *prima facie* sense of diversity is more attainable. Maintaining the exact ratio of treatment to control schools of 16 to 9 requires at least 25 schools per district, or four districts altogether. Four districts, however, does not provide much scope for observing impacts in urban and rural settings, in different states, or in different ethnic or language communities. An approximate ratio can still be achieved with some districts having 8 treatment and 5 controls or 8 treatment schools and 4 controls. This means spreading the sample over eight districts, which provides more scope for conducting the study in a variety

of settings in terms of different regions of the country and possibly different student body ethnic compositions.

### 4. Power for Subgroup Analysis

The statistical power for subgroup analysis can be estimated by assuming a smaller number of students per school. Our calculations suggest that statistical power does not decline noticeably until there are fewer than 10 students per school (Figure III.1). Starting with samples of about 30 students per school, we can construct subgroups with as few as one-third of the full sample of students and still be able to detect impacts of about 0.26 standard deviations in a sample of 100 schools.

## C. Data Collection

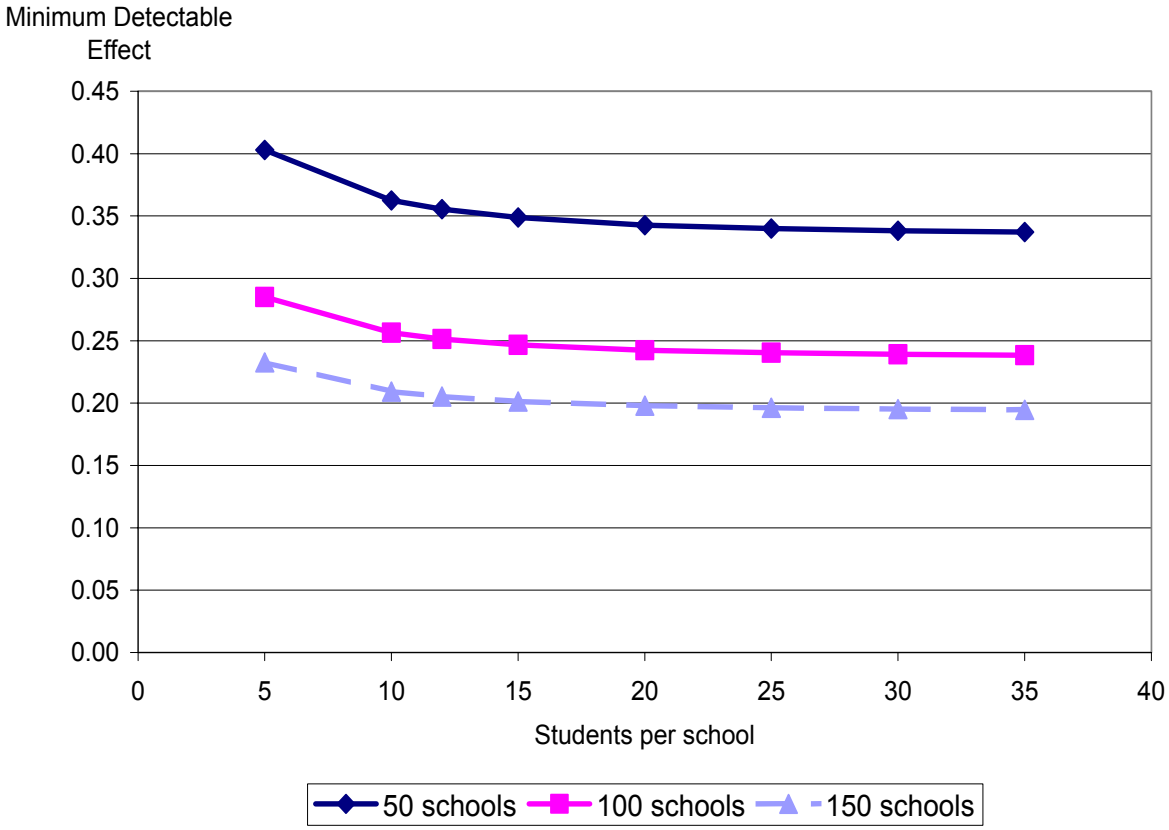
The data required for the random assignment design described here would cover student achievement; other characteristics of students, teachers, and schools; and information on the presence of various school conditions and practices. Some of the data items can be collected through quantitative methods, including experimenter-administered tests, survey instruments, and school records. Other data items may need to be collected through qualitative methods, such as site visits. Information collected at baseline, i.e., before the interventions are implemented, would be particularly useful for isolating the effect of interventions during the study's observation period, allowing us to remove prior background knowledge and other pre-baseline influences from the analysis. Information collected at a followup point, typically at the end of the school year, would be used to assess the interventions' effect on outcomes.

Adding a followup point at one year, two years, or both points after the end of the implementation year would provide evidence on whether the impacts observed at the end of the implementation year persisted or faded. It may be desirable to condition Year 2 and Year 3 followups on the presence of a positive impact at the end of Year 1 of the study. Therefore, a multi-intervention study may lead to conducting followup on a subset of interventions.

The issue of how to assess learning outcomes targeted by reading comprehension interventions arose frequently during TWG meetings. There was a strong consensus that the main outcomes should be students' ability to extract meaning from text, i.e. to read for understanding in content areas like sciences and social studies. There was also a recognition that the newly developed skills should help schools both meet their immediate AYP goals according to federal standards and raise scores on state assessments, which are playing an increasingly important role in accountability systems. The alignment, however, of state assessments, regular district-administered tests, and the learning outcomes that relate to extracting meaning from text is tenuous at best. As a result, the study designers face a number of choices and tradeoffs about how to measure outcomes and define success for the intervention.

FIGURE III.1

MDE BY NUMBER OF STUDENTS PER SCHOOL




- ***State Assessments.*** State standards are mentioned explicitly in the No Child Left Behind legislation, which implies that state assessments should be the preferred yardstick for student performance. Besides being of great interest to policymakers, data from state assessments may be inexpensive to obtain for the study if they are already being administered in the grades and subjects of interest. However, the states where the study schools are located may not administer tests that target the right grades and subjects. Furthermore, comparisons of assessment data across states could be problematic, given the state-to-state variation in the assessments. Also, State assessments may be limited in the extent to which they can provide baseline data for the study.

- ***District-Administered Tests.*** Many school districts typically administer standardized tests in the spring in addition to or instead of state-mandated assessments. Like state assessments, data from district-administered tests are

relatively inexpensive to obtain but would not necessarily be comparable across geographic areas and or provide baseline data in the fall. It is possible, but not necessarily likely, that districts from different geographic areas selected for the study would be using the same tests from the same publisher.

- ***Evaluator-Administered Tailored Assessments.*** The ideal assessment might be one that the evaluator could both tailor to the specific cognitive skills targeted by the interventions and administer in the fall and spring. The drawback to this option, in addition to the cost of administering a new test, would be the difficulty of developing a new test. Furthermore, a highly tailored test would have to sacrifice breadth of content domains in order to measure depth of achievement in very specific skill areas such as reading for understanding, vocabulary, and acquisition of background knowledge in specific content areas.

- ***Evaluator-Administered Standardized Test.*** The evaluation team could administer its own test, but by selecting an off-the-shelf instrument could avoid the burden of a test-development period. This approach still imposes the burden of administering a separate test in addition to those routinely given in the study schools but offers the following advantages: it would be administered uniformly across study sites, it gives the evaluation team some flexibility to choose test length and coverage, and probably most important, it provides comparable measures across study sites that can be easily interpreted or, in other words, compared to national norms.

MPR recommends a hybrid data collection approach based heavily on evaluator-administered standardized tests but involving district-administered tests, including student performance on state assessments, as well. Some of the district and state test information will have to be used for district-specific or state-specific analyses, but it can provide a useful check on the robustness of the findings derived from the evaluator-administered tests.

This recommendation is practical from an evaluation design perspective and is consistent with the advice received from many experts we consulted. One TWG panelist, who is also a member of the IRP, wrote in a commissioned paper, "The pre- and post-treatment measures should include standardized reading comprehension tests (e.g., Woodcock-Johnson) and school or district assessments of reading proficiency. Since the instruction is to be carried out in the content areas of science or social studies, then knowledge of subject matter tests could be developed that focused on key concepts and relationships" (Trabasso 2003).

Other achievement data in addition to test scores can be collected to round out the picture of achievement. These data could include student grades, retention in grade, attendance, and placement into special education or remedial reading programs. Such information is commonly available in school records and would be relatively inexpensive to obtain.

Assessing fidelity of implementation is also important to interpreting impacts. Because the study focuses on implementing particular reading comprehension interventions in districts, schools, and with teachers that have not yet used them, some implementation efforts may not succeed within the study's timeframe. The design team recommends creating metrics to assess the fidelity of implementation, possibly based on developer guidelines and other sources, so that measures of implementation success can be applied as part of the effort to understand impacts. The national study would provide useful information to the field if it was found that some interventions could produce effects even in the face of implementation difficulties or when they appeared to fall short of developer guidelines while others did not produce effects even when implemented with great fidelity to the ideal.

In addition, the opportunity to assess fidelity of implementation could be used to collect data on various instructional and environmental components of interest. Trained field researchers could visit schools and classrooms participating in the study to observe implementation and to gather information about conditions and practices through qualitative techniques such as interviews with teachers and administrators. The researchers could code the information into variables that would be used in the hierarchical linear modeling (described below) to measure the relationship between the conditions and practices (indeed, fidelity of implementation can be viewed as a practice) and outcomes. Field researchers could also visit schools or classrooms in the control groups if a no-intervention control is used, helping to build a better understanding of the extent to which comprehension instruction is provided in the absence of a formal intervention. A fuller protocol for field research needs to be considered, together with quantitative instruments, so that the different types of information will be mutually supportive.

## D. ANALYSIS

Experimental designs yield a simple estimator of the effects of an intervention: the difference between the average outcomes of treatment and control groups at followup (or equivalently, the differences in mean outcomes between groups that received different interventions). The analysis plan for the effective practices study as described here should build on this simple estimator in a few important ways in order to properly address the three research questions. Consistent with our power analysis explained in Section B, we recommend that the analysis plan for impact estimation and hypothesis testing follow some simple guidelines:

- ***Use Regression Adjustment to Increase Precision.*** As in most studies of student achievement, baseline test scores can increase precision a great deal by explaining most of the variation that existed before the students were exposed to interventions. Other background variables may also be used in a regression framework as long as they are not influenced by assignment to treatment status. We recommend regression adjustment based on as many of these variables as possible. School- and community-level background variables will be especially useful to include, because they help explain variation in the school-level regression model, the same level of analysis at which treatment status varies.

- ***Account for the Nesting of Students Within Classrooms and Schools.*** It is common in education settings that students are nested within classrooms that are nested within schools. The fact that students within classrooms share a set of classroom characteristics—i.e., they have the same teacher and they learn together—means that their performance is not independent of one another. Similarly, classrooms in the same school share an instructional leader (principal), and the teachers typically interact. Ignoring these dependent relationships usually results in an artificially low estimate of the standard error, i.e. the uncertainty around the impact estimates. The conventional method for dealing with this type of non-independence of observations is to estimate a hierarchical linear model (HLM) of student achievement. We recommend that all impact estimation use HLM or an equivalent method to ensure that the correct standard errors are used for hypothesis testing.

- ***Account for the Multiple Comparison Problem.*** As discussed in Section B, a multi-interventions experiment can lead to faulty inferences if one is not careful. In the face of multiple treatments, it is important to recognize that as the number of comparisons goes up, the probability of finding an effect that is "statistically significant" when no real difference exists goes up also. We therefore recommend that all analyses use the proper techniques for controlling what is known as the Type I error rate, the probability of falsely detecting a meaningful difference. These technique were discussed in Section B above, which reported on the power analysis.

- ***Propose Subgroup Analysis in Advance.*** Subgroup analysis in the context of the proposed study is straightforward. All of the methods can usually be applied to the subsample represented by members of a particular subgroup of interest. Subgroups can be defined in terms of student characteristics such as family income, limited English proficiency, or pre-test score on a test of word-level skills. They can also be defined in terms of school or district characteristics, such as school size or urbanicity. The only serious pitfalls are defining subgroups in terms of characteristics that can be influenced by treatment status (such as quality of curriculum implementation) or defining subgroups after the data are observed. Subgroup analyses should usually be justified on theoretical grounds *a priori* to avoid the appearance of improper data-mining.

*PAGE IS INTENTIONALLY LEFT BLANK TO ALLOW FOR DOUBLE-SIDED COPYING*

# REFERENCES


Barr, Rebecca. "Interventions for Children Experiencing Early Reading Difficulties." In *Successful Reading Instruction*, edited by Michael L. Kamil, JoAnn B. Manning, and Herbert J. Walberg. Greenwich, CT: Information Age Publishing, 2002.

Block, Cathy Collins, and Michael Pressley, Eds. *Comprehension Instruction: Research-Based Best Practices.* New York: Guilford Press, 2002.

Box, George P., William G. Hunter, and J. Stuart Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* New York: John Wiley, 1978.

Cook, Thomas, and Donald Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Hopewell, NJ: Houghton Mifflin, 1979.

Duke, Nell K. "For the Rich it's Richer: Print Environments and Experiences Offered to First-grade Students in Very Low- and Very High-SES School Districts." *American Education Research Journal* vol. 37, no. 3, 2000, pp. 456-457.

Gersten, Russell, Lynn S. Fuchs, Joanna P. Williams, and Scott Baker. "Teaching Reading Comprehension to Students With Learning Disabilities: A Review of the Research" *Review of Reading Research*, vol. 71, number 2, Summer 2001, pp. 279-320.

Guthrie, J.T. "Engagement and Motivation in Reading Instruction." In *Successful Reading Instruction*, edited by Michael L. Kamil, JoAnn B. Manning, and Herbert J. Walberg. Greenwich, CT: Information Age Publishing, 2002.

Kamil, Michael. "Desirable Characteristics of Comprehension Interventions." Paper commissioned by the Institute of Education Sciences. December 2003.

Myers, David, and Mark Dynarski. *Random Assignment in Program Evaluation and Intervention Research: Questions and Answers.* Princeton, NJ: Mathematica Policy Research, June 2003.

National Reading Panel. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction.* Washington, DC: National Institute of Child Health and Human Development, December 2000.

Pressley, Michael. "Comprehension Strategies Instruction: A Twentieth Century Report" In Block and Pressley, Eds. *Comprehension Instruction: Research-Based Best Practices.* New York: Guilford Press, 2002.

Trabasso, Tom. "Teaching of Comprehension in Content Areas: Design Criteria." Paper commissioned by the Institute of Education Sciences. December 2003.

# APPENDIX A

# TWG AND IRP MEMBERS

---

## TECHNICAL WORKING GROUP MEMBERS

Thomas Cook, Northwestern University

Jack Fletcher, University of Texas-Houston

David Francis, University of Houston

John Guthrie, University of Maryland-College Park

Robinson Hollister, Swarthmore College

Stephen Raudenbush, University of Michigan-Ann Arbor

Catherine Snow, Harvard University

Joseph Torgeson, Florida State University

Thomas Trabasso, University of Chicago

**INDEPENDENT REVIEW PANEL MEMBERS**

Kaleem Caire, American Education Reform Council

Thomas Cook, Northwestern University

Christopher Cross, Center on Education Policy

Gayle Fallon, Houston Federation of Teachers

David Francis, University of Houston

Norma Garza, Rodriguez, Colvin, & Chaney

Eric Hanushek, Hoover Institution

Sharon Johnson, Withrow University High School

Paul Peterson, Harvard University

Stephen Raudenbush, University of Michigan-Ann Arbor

Eric Smith, Anne Arundel (Maryland) Schools

John Stevens, Texas Business and Education Coalition

Patricia Supple, Archdiocese of Los Angeles

Tasha Tillman, Friends of Choice in Urban Schools

Thomas Trabasso, University of Chicago

Maris Vinovskis, University of Michigan-Ann Arbor

Rodney Watson, Louisiana Department of Education

# APPENDIX B

# TEACHING OF COMPREHENSION IN THE CONTENT AREAS: DESIGN CRITERIA

by Thomas Trabasso, University of Chicago

December, 2003

## Teaching of Comprehension In Content Areas

Reading comprehension has been defined as the “intentional thinking during which meaning is constructed through interactions between text and reader” (Durkin, 1993). According to this view, meaning resides in the intentional, problem-solving, thinking processes of the reader that occur during an interchange with a text. The content of meaning is influenced by the text and by the reader's prior knowledge and experience that is brought to bear on it. Reading comprehension is the construction of the meaning of a written text through a reciprocal interchange of ideas between the reader and the message in a particular text (Harris and Hodges, 1995, defintion # 2, p. 39). The bulk of instruction of text comprehension research during the past two decades has guided by this cognitive conceptualization of reading.

A reader reads a text in order to understand what is read and to put this understanding to use. A reader can read a text in order to learn, to find out information, or to be entertained. These various purposes of understanding require that the reader use knowledge of the world, including language and print. This knowledge enables the reader to make meaning of the text, to form memory representations of these meanings, and to use them to communicate with others information about what was read.

In terms of schooling, it is in the third grade that readers need to perform acts that lead to good comprehension, learning, and memory. The focus on comprehension instruction should begin in the third grade, though it is possible to introduce comprehension instruction (answering questions) in earlier grades.

### Goals

The central goals of teaching comprehension are to develop procedures in reading that enable the reader to read texts in content areas at grade level and be able to use what they have read. These uses include paraphrasing what was read in their own words, summarizing the text accurately and including the central ideas of the text, posing and answering self-generated or teacher generated questions about the content, and remembering what was read by recalling the content accurately after reading with different delays.

These procedures are necessary to the learning of content in subject matter areas and should be taught within them on content relevant texts within a curriculum rather than in isolation on arbitrary texts.

### Basis for Adoption of a Form of Comprehension Instruction

There are three reasons for choosing to teach a particular comprehension strategy: (1) it has a proven scientific basis as to its effectiveness, (2) it can be readily taught to teachers, and (3) teachers can use it flexibly in conjunction with other strategies during reading of a text.

### Areas of Comprehension Instruction

Three areas of instruction that should be addressed in the research are (1) instruction of vocabulary, (2) instruction of text comprehension, and (3) preparing classroom teachers to teach comprehension.

There are a variety of methods where readers acquire vocabulary through explicit instruction and improve their comprehension of what they read. Further, there has been considerable success in teaching a variety of effective text comprehension strategies that lead to improved text comprehension. The most promising lines of research within the reading comprehension strategies area have focused on teacher preparation to teach comprehension.

Teachers can be helped by intensive preparation in strategy instruction and this preparation leads to improvement in the performance of their students.

Instruction of vocabulary should not be separate from direct instruction of text comprehension during reading of content area texts. In content areas, special vocabulary is required to understand the area's concepts and relationships. This is particularly true of science but also is true of social studies. However, these vocabularies carry general meaning as well so that learning them in context is not restrictive.

The learning of vocabulary is best carried out in the context of reading and trying to understand the content area texts. Comprehension strategies such as thinking aloud, question generation, and question answering by the reader or teacher during reading should reveal to the teacher where the reader is experiencing difficulty in understanding particular concepts or relationships. These difficulties present the teacher with opportunities to teach definitions of concepts and their relations or to refer the reader to sections of the text that define concepts and relate them.

The idea behind explicit instruction of text comprehension is that comprehension can be improved by teaching students to use specific cognitive strategies or to reason strategically when they encounter barriers to comprehension when reading. The goal of such training is the achievement of competent and self-regulated reading. Instruction in comprehension strategies is carried out by a classroom teacher who demonstrates, models, or guides the reader on their acquisition and use. When these procedures are acquired by the reader, the reader becomes independent of the teacher. Using them, the reader can effectively interact with the text without assistance. Readers who are not explicitly taught these procedures are unlikely to learn, develop, or use them spontaneously.

There are eight kinds of comprehension instruction that appear to be effective and most promising for classroom instruction. These are, in alphabetical order, (1) comprehension monitoring (the reader learns how to be aware or conscious of his or her understanding during reading and learns procedures to deal with problems in understanding as they arise),(2) cooperative learning (readers work together to learn strategies in the context of reading),(3) graphic and semantic organizers (the reader writes or draws graphically the meanings and relationships of the ideas that underlie the words in the text, (4) story structure (the reader learns to ask and answer who, what, where, when, and why questions about the plot and, in some cases, maps out the time line , characters, and events in stories),(5) question answering (the reader answers questions posed by the teacher and is given feedback on the correctness), (6) question generation (the reader asks himself or herself what, when, where, why, what will happen, how and who questions), (7)
summarization (the reader attempts to identify and write the main or most important ideas that integrate or unite the other ideas or meanings of the text into a coherent whole), and (8) multiple strategy teaching (the reader uses several of the preceding procedures in interaction with the teacher over the text).

Multiple strategy teaching typically includes self-generated questioning and answering, summarizing, and prediction (#5, 6, and 7 above). It is effective when these procedures are used flexibly and appropriately by the reader or the teacher in naturalistic, learning contexts. Multiple strategy instruction that is flexible as to which strategies are used and when they are taught over the course of a reading session provide a natural basis upon which teachers and readers can interact over texts.

The important development of instruction of comprehension research is the study of teacher preparation for instruction of multiple, flexible strategies with readers in natural settings and content areas and the assessment of the effectiveness of this instruction by trained teachers on comprehension. Acquiring and practicing strategies in isolation and then attempting to provide transfer opportunities during the reading of text is not the kind of instruction that is required in naturalistic contexts. Proficient reading involves a constant, ongoing adaptation of many cognitive processes. Thus, teachers must be skillful in their instruction and must respond flexibly and opportunistically to students' needs for instructive feedback as they read. In order to be able to

do this, teachers must themselves have a firm grasp not only of the strategies that they are teaching the children but also of instructional strategies that they can employ to achieve their goal. Many teachers find this type of teaching a challenge, most likely because they have not been trained to do such teaching.

There have been two major approaches to comprehension strategy instruction in the classroom: Direct Explanation and Transactional Strategy Instruction. In the Direct Explanation approach, teachers do not teach individual strategies but focus instead on helping students to (a) view reading as a problem-solving task that necessitates the use of strategic thinking, and (b) learn to think strategically about solving reading comprehension problems. The focus is on developing teachers' ability to explain the reasoning and mental processes involved in successful reading comprehension in an explicit manner, hence the use of the term "direct explanation". The implementation of direct explanation requires specific and intensive teacher training on how to teach the traditional reading comprehension skills found in basal readers as strategies, e.g., to teach students the skill of how to find the main idea by casting it as a problem-solving task and reasoning about it strategically.

The Transactional Strategy Instruction approach includes the same key elements as the Direct Explanation approach, but it takes a somewhat different view of the role of the teacher in strategy instruction. The Transactional Strategy Instruction approach focuses on the ability of teachers to facilitate discussions in which students (a) collaborate to form joint interpretations of text, and (b) explicitly discuss the mental processes and cognitive strategies that are involved in comprehension. In other words, the emphasis is on the interactive exchange among learners in the classroom, hence use of the term "transactional."

In both approaches, teachers explain specific strategies to students and model the reasoning associated with their use. Both approaches include the use of systematic practice of new skills, as well as scaffolded support, in which teachers gradually withdraw the amount of assistance they offer to students. The different emphases of the two approaches (explanation vs. discussion) result in differences in the level of collaboration among students.

Effective teachers need training to explain fully what it is that they are teaching (what to do, why, how and when), to model their own thinking processes for their students, to encourage students to ask questions and discuss possible answers and problem solutions among themselves, and to keep students engaged in their reading by providing tasks that demand active involvement.

In sum, teachers should be trained in flexible, multiple strategy, comprehension instruction where they interact with students during the reading of a text in a subject matter field such as social studies or science.

Content Areas

Comprehension should be taught in content areas where expository or narrative texts are used. The two content areas are social studies and science. The content areas provide rich contexts in which concepts (words) and relationships are learned. These content areas are likely to be a part of the curriculum in grades 3 through 6. If these areas are taught to children in grades 1 and 2, then comprehension procedures may be introduced at this level, though less is known here. Grades 3-6 have been the main focus of comprehension instruction research.

Grade Levels

Research on comprehension instruction has been most successful in grades 3 through 6, with greater success achieved in grades 5 and 6. Grades 3 and 4 are the ones where readers who have acquired decoding and fluency skills must now use those skills to acquire information from reading the text. Comprehension instruction in content areas for grades 3 and 4 are likely to benefit both reading and learning. It is important that readers in grades 3 and 4 be assessed for their ability to decode and to read fluently since these abilities are necessary for comprehension.

Reading Levels

Research on comprehension instruction has been shown to be beneficial to readers who are reading below grade level. In a study of comprehension instruction intervention, the reading level of the participants in the experimental and control groups should be obtained as a part of the pre-treatment phase of the study. Reading scores from standardized reading tests such as the Woodcock-Johnson and from state or district design reading assessments should be obtained. These same measures could be used in post-treatment assessment.

The pre-treatment reading assessments should be used to define levels of reading competence so that comprehension instruction by reading level effects can be determined in the research.

Kind of Comprehension Instruction

The main method of teaching vocabulary in context is an association method where learners are encouraged to draw connections between what they do know and words they encounter that they do not know during reading. Explicit Instruction can be given by teachers when the reader indicates a lack of knowledge about a word. This can be accomplished by giving definitions or other attributes of the words appropriate to the context and content that is being learned. Implicit Instruction can also occur where students are exposed to words or given opportunities to do a great deal of reading in the content area. Here the definitions and examples in the text can assist in the learning of vocabulary. Texts should be examined for the degree to which they explicitly define, illustrate, and provide visual and other support for the meaning of words. The teaching of content area matter typically uses multi-media methods. Here, content area vocabulary is used and taught in discussion or in other modes such as accompanying pictures in the text or in the use of film, dramatizations, or direct experience in viewing concepts and relations that is referred to in the text. It is well known that repeated exposures to vocabulary items is an important experience for learning gains. The best gains are made in instruction that extends beyond single class periods and involves multiple exposures in authentic contexts. Instruction of vocabulary words by the teacher prior to reading can also facilitate both vocabulary acquisition and comprehension. Dependence on a single kind of vocabulary instruction method will not result in optimal learning. A variety of methods should be used with an emphasis on multimedia aspects of learning, richness of context in which words are to be learned, and the number of exposures to words that learners receive.

The kind of instruction should be direct. The focus of the instruction should be on teacher-reader(s) interaction and teacher/peer/group demonstration and practice of flexible cognitive strategies that readers learn to employ during reading of texts in content areas. These learning and use of multiple strategies flexibly should promote self-awareness of cognitive processes that occur during reading. The most common form is to have a teacher demonstrate or model for readers the action(s) that readers then perform themselves such as asking questions and trying to answer them, paraphrasing a sentence or paragraph, summarizing, and predicting what will occur next. A peer or group of peers can perform the same function of the teacher in collaboration with a reader. Readers can demonstrate and practice strategies collaboratively. Readers need to practice strategies with assistance until they achieve a gradual internalization and independent mastery of them. The emphasis should be on the flexible use of different strategies by the teacher and reader, though the teaching of one strategy at a time may be necessary in some cases to learn a new strategy. Skilled reading involves an ongoing adaptation of multiple cognitive processes. Becoming an independent, self-regulated thinking reader is the product of years of development.

The major problem facing the teaching of reading comprehension strategies is that of implementation in the classroom by teachers in a natural reading context with readers of various levels on reading materials in content areas. For teachers, the art of direct instruction involves a series of "wh" questions: knowing *when* to apply *what* strategy with *which* particular student(s). Having students actually develop independent, integrated strategic reading abilities may require subtle instructional distinctions that go well beyond techniques such as direct instruction, direct explanation, or reciprocal teaching (Duffy, 1993). Duffy argues that strategies are not skills that

can be taught by drill but plans for constructing meaning. Teaching students to acquire and use strategies may require altering traditional approaches to strategy instruction. It may be necessary to free teachers of the expectation that their job is to follow directions narrowly. Being strategic is much more than knowing the individual strategies. When faced with a comprehension problem, a good strategy user will coordinate strategies and shift strategies as it is appropriate to do so. They will constantly alter, adjust, modify, and test until they construct meaning and the problem is solved.

Texts and Content Areas

The teaching of comprehension in isolation from content areas where understanding and learning co-occur. Text, as a variable, has been sorely neglected. The external validity of a study could also be improved by the kind of texts used (both expository and narrative and sampled from content areas), analysis of text difficulty, the content and structure of the text, the vocabulary and sentence complexity of the text, appropriateness of the level of text difficulty to the ability of readers, and possible interactions between difficulty of the text and ability of reader. Long-term benefits could be assessed through follow up studies later in time so that the effects are not just short term.

Time on Task

It is unclear how many hours of instruction should be devoted to the teaching of comprehension in a content area. In experiments, 20 hours seems to be a minimal amount of time necessary to achieve results with below grade level readers. Since the intervention is embedded within the social science or science curriculum, it might be possible to devote 15 minutes to a half hour of comprehension instruction on texts that are being read as course requirements. This instruction would constitute a more intensive form of learning the content via reading the text. The range of instruction in studies has been from 20 to 100 hours but it is not clear how much instruction is optimal nor whether extending the number of hours adds more benefit since these comparisons have not be made on existing studies. Perhaps a module of instruction within the content area (e.g., six weeks) would serve as the basis for the duration of the supplementary comprehension instruction. The modules would depend upon the nature of the curriculum for science and social studies. It might be desirable to teach comprehension on more than one module and to assess transfer to subsequent modules where the reader is not trained.

Assessments, Measures, and Experimental Design

The experimental design would employ a Pre-Post design with a treatment and control groups. Measurements would be taken on comprehension and use of information read in the modules where instruction occurs and in transfer to new modules. The treatment groups would receive comprehension instruction as outlined above and the controls would have some kind of teacher-reader interaction over text. In the latter, teachers are not trained but interact as they normally would in instruction over a text with the reader(s). The interaction is to assure active involvement by both experimental and control participants. Random assignment to experimental and control groups would be achieved at the classroom level.

The pre- and post-treatment measures should include standardized reading comprehension tests (e.g., Woodcock-Johnson) and school or district assessments of reading proficiency. Since the instruction is to be carried out in the content areas of science or social studies, then knowledge of subject matter tests could be developed that focused on key concepts

and relationships. These knowledge assessments would serve two purposes: (1) the pre-test is a measure of knowledge in the subject matter field before instruction and (2) the post-test is a measure of learning gains achieved from the course and comprehension instruction. The gains achieved by treatment versus control groups could be compared to assess effects of instruction on comprehension over and above those of the course itself. Further, the knowledge assessments in pre-testing provide a basis for the study of individual differences in vocabulary and knowledge of the content area prior to learning and comprehension instruction. It would allow for an analysis of the knowledge level by treatment interaction where benefits to low knowledge readers could be assessed compared to those for high knowledge readers.

Other outcome measures would include those used by teachers to assess student performance over the course of study (e.g., assignments, workbooks, etc. ) as well as grades. It would be important for the classrooms in the treatment and control conditions to follow common curriculums, including lessons and texts so that what is being learned is common to both conditions.

### Teacher Training (Professional Development)

Procedures for the professional development and in-service training of teachers are necessary for teaching reading comprehension in a subject matter area. Subject matter teachers would have to be trained in workshops on the methods that they will use in comprehension instruction. This training might be accomplished during the summer prior to implementation in the classroom. Studies relevant to teaching teachers on teaching of reading comprehension have been done by Gerald Duffy and Michael Pressley. The experiences of these researchers would inform the development and implementation of teacher training.

### Analyses of Texts

The texts that will be used in comprehension instruction are to be sampled from the books that are in actual use in the content area curriculum. The texts should be representative of the content being taught and correspond to a lesson unit. The text should be analyzed for grade level appropriateness. The Lexile method could be used since it yields quantitative, grade-level scores, and is sensitive to word frequency and sentence length. The length of text should correspond to that used in a normal lesson within the content area. The concepts and relations of the text should be analyzed and used in vocabulary teaching during reading. Identification of explicit definitions versus implicit definitions of concepts should be made in the texts since explicit definitions facilitate understanding, especially for poorer readers.

For readers reading below grade level, it is recommended that a text that corresponds to their level be used instead of one above their level. The reason for this recommendation is that one can have more success in comprehension instruction for an appropriate grade level text than one that is above level.

### Fidelity to Treatment

Audio-Video recordings of both the teacher and the reader(s) should be made during comprehension instruction. These recordings are to be used to analyze what both the teacher and reader does, assessing fidelity to treatment by the teachers and the effectiveness of the instruction on student procedures during reading. Fidelity to treatment is necessary to insure that teachers taught the way that they were trained. Variation in teacher methods could also be analyzed and validated against student performance.

Further, what students do can be analyzed over the course of instruction to examine learning and transfer.

# APPENDIX C

# DESIRABLE CHARACTERISTICS OF COMPREHENSION INTERVENTIONS

by Michael Kamil, Stanford University

December, 2003

The following is a skeleton outline for the paper that follows.

Reader Variables

I. General background factors
   World knowledge
   Cultural issues
   Parental/family variables
II. Oral language components
   Phonemic awareness
   Proficiency
      Vocabulary
      Auding ability
III. Reading components
   Phonics and word identification
   Reading vocabulary
   Comprehension strategies
IV. Student motivation
   Motivation and engagement

Task Variables

V. Application
   Setting goals or purposes
   Writing
VI. Practice
   Fluency
   Reading and writing

Text Variables

VII. Texts
   Genre
   Complexity
   Interest
   Electronic and multimedia texts

Teacher Variables

VIII. Teacher Professional Development
   Facility in delivering the intervention

VIII. Cost-Benefit analysis
   Close the gap
   Ease of implementation
   Relative cost
X. Technology
   Instructional uses of computer (and other) technology

## Desirable Characteristics of Comprehension Interventions

### General Considerations

Comprehension is defined as reading comprehension as the process of simultaneously extracting and constructing meaning through interaction and involvement with written language (RAND Reading Study Group, 2002, p. xiii). The assumption in the following is that this discussion assumes that there are students who have difficulty with comprehension. However, most instructional strategies will not be specific to students who struggle with comprehension. Rather, most strategies will improve comprehension for most students.

There are four components which must be addressed in any discussion of comprehension: Reader, Text, Task, and Context. A discussion of comprehension *instruction* requires the addition of a teacher or other delivery system as an additional component. Teachers must deliver the instruction, and teacher quality is an issue in that it is clearly related to student performance. At other times, the delivery system could be computer with appropriate software.

In general, any intervention must accomplish several general goals. It should close the gap between good and poor comprehenders, even if it improves comprehension for all readers. Thus, it should not, in general, accentuate potential or real Matthew effects often observed as the result of educational interventions. While this is a difficult goal to achieve, it should be the aim of all interventions. In addition, interventions should be cost-effective. That is, student outcomes should be commensurate with the resources required to implement and sustain the intervention. Implementations with higher costs in resources should generate greater improvements. This suggests that interventions, in general, should not be highly complex or require a great deal of specialized professional development. If they are, however, the expectations of much more substantial results would be appropriate.

Interventions must also contain an appropriate, comprehensive assessment system. The assessment system should be such that it is both instructionally formative and summative. That is, it should allow decisions about when students reach various proficiency benchmarks. It should also provide information about what is needed instructionally when students do NOT reach those benchmarks.

The assessment system in an intervention is more than an “add-on”. It should be the crucial element in decision-making with regard to instruction. It needs to be integral to the intervention so that users can monitor specific learning outcomes rather than more general outcomes available from, for example, off-the-shelf standardized assessments.

In what follows, each of the elements in the outline above is elaborated to place them in context, suggesting what sorts of importance each should assume in an ideal intervention. However, the attempt to generate a description of a single set of criteria for

an intervention is foolhardy. There are different needs for different students at different ages.

## Oral Language and Comprehension

In general, reading is taught in the early grades as an oral activity. The "theory" is that students can comprehend reasonably complex oral language. If they could be taught to translate (decode) text to speech they could then monitor the decoded text as if it were speech and comprehend it. There is an essential for this to work: The text must be a "reasonable" match to the student's oral language. That is, if the text were read to the student, the student should be able to comprehend it. At least a large portion of this "match" falls on oral vocabulary. If the student decodes a word to speech, it must be in the reader's oral vocabulary for it to be meaningful. For example, even though a student might decode a nonsense word like "ferple", there would be no way to assign a meaning to it. Ferple would remain incomprehensible even if it were decoded properly.

Oral language comprehension has been termed auding by Sticht and his colleagues. They showed that up to approximately grade 3, oral language comprehension and reading comprehension are roughly interchangeable. That is, improvements in auding lead to improvements in reading, while improvements in reading lead to improvements in auding.

Thus, one of the primary requirements in an intervention to improve comprehension must be the attention to oral language abilities. Assessment should monitor oral language abilities for diagnostic purposes as well as progress monitoring.

There are additional factors that affect comprehension, including general background and cultural knowledge. These factors can be conceived of as part of oral language or as more general abilities. In either case, any intervention should deal with these issues, either in school or as outreach to parents. (A good example of the ability to work with parents to improve children's literacy outcomes is the work of Neuman with single mothers and literacy objects, even if this was with very young children.)

## Phonics and Word Level Processes

While many believe these word level processes are not implicated in comprehension, the data seem clear that they at least set the preconditions for skilled comprehension. Curtis (In Press) goes so far to maintain that as many as 10% of adolescent readers who have problems can be helped by some type of remediation directed at word identification. Even though this is not a generalized basis for a comprehension intervention, any intervention should, at the very least, screen for these issues. The intervention should provide appropriate instructional strategies for those students in need of remediation.

## Comprehension Instruction

Vocabulary

There is much evidence that vocabulary is a strong determiner of comprehension. Research (e.g. Anderson & Freebody, 1984; Stanovich, Cunningham, & Freeman, 1984) has shown that reading ability and vocabulary size are related, but the causal link between increasing vocabulary and an increase in reading ability has been difficult to demonstrate (Stanovich, 2000, p. 162).

Anderson, Wilson, & Fielding (1988) showed large differences in amounts of daily reading among children. The number of words encountered in leisure reading per year at leisure varied from 8 to 4,700,000. These enormous variations in reading, of course, lead to large differences in children's' vocabularies and comprehension abilities. Hart & Risely (1995) report similar findings, but identified these deficits in at-risk, low SES students.

Comprehension interventions should at least have a strong focus on vocabulary and vocabulary development, where necessary. Once again, an embedded assessment system should screen for vocabulary and provide appropriate instructional strategies where appropriate.

Strategy Instruction

Any comprehension intervention should provide instruction in comprehension. One of the areas of comprehension instruction is to assist students to learn comprehension strategies. There are two major formats for comprehension strategy instruction, direct or explicit instruction and transactional instruction. Direct instruction tends to be teacher dominated while transactional instruction is more like to be student centered. There is probably a case to be made for some combination of both approaches, depending on student needs.

For strategy instruction, there is also the decision about whether to teach students single strategies or multiple strategies. Evidence is accumulating that suggests there is a real benefit to the multiple strategy approach. As this becomes clear, it would also seem that this should be a desired component for comprehension interventions.

## Text

Genres

Texts are a critical variable in comprehension interventions. As noted early in this paper, there are (at least) two distinct types of text—generally called literary or story texts and information or expository. There are many other genres, but this is an accepted distinction by many researchers and practitioners. Primary grade reading is grounded primarily in story genres; upper grade reading is done primarily in information genres. There is at least the occasional suggestion that the fourth grade slump can be attributed to

this mismatch between the texts of early reading instruction and the text that must be read in later grades.

Any intervention should clearly attend to the genre issue, both providing appropriate instruction in reading a wide variety of genres, as needed. At the same time, students should be given ample opportunity to practice reading in the variety of genres, in addition to the instruction.

One other genre that needs attention is multimedia text. Multimedia text is text that combines both conventional text and other visual media to convey information. This is an important issue for interventions because textbooks are currently being produced to include far more graphical information. Layouts are much more varied than conventional textbooks were. (In fact, many of these new texts look much like web pages.) The importance of multimedia text for interventions is that this may be a locus of difficulties in comprehension. Students may not know how or why to combine multimedia information with the conventional text. Other students may not be able to access the conventional text and rely ONLY on the multimedia elements. In either case, intervention is important because information is, increasingly, being presented in these mixed formats.

Equally important is that text should be at appropriate levels of difficulty, both in terms of readability as well as conceptual complexity. Instruction in a comprehension intervention should vary appropriately with the types of text being used.

## Motivation

Motivation (in reading) can be defined as the cluster of personal goals, values, and beliefs with regard to the topics, processes, and outcomes of reading that an individual possesses (Guthrie and Wigfield, 2000, p. 404). This is not the same thing as interest, attitude, or beliefs (Guthrie and Wigfield, 2000). One could have an interest in reading, but nevertheless choose not to read. Motivation is the underlying factor that disposes one to read or not. Engagement is yet another variable in this affective cluster of concepts. Engagement in reading is the extent to which an individual reads to the exclusion of other activities, particularly when faced with the other choices.

Any intervention MUST account for students who can read but choose not to as well as those who do not read because they cannot. Both types of students need to have a reason to exert the effort required to learn to read. These considerations are related to some of the considerations in the next section on tasks.

## Tasks

Of particular importance in comprehension interventions is the type of task that the reader is being asked to do as a result of reading the text. Much comprehension instruction is done in impoverished contexts, with the student reading a passage and

answering some questions about it. Independent reading depends on the student being able to set purposes and goals, and adapt reading to them, as a function of the text and the reason for reading it. Struggling readers may not be able to understand how to set purposes for different tasks. Interventions should stress the differences among tasks and the consequences that different tasks have for reading and comprehending.

Many of the tasks following reading involve writing, and, consequently, writing should be an integral part of any comprehension intervention. Writing should include the same emphasis on diverse genres recommended for reading instruction above. There should be relatively explicit instruction in connecting reading and writing to promote comprehension.

## Practice

Related to tasks is the notion of practice. While the National Reading Panel found no experimental evidence for the notion that reading practice improves reading, it seems to many like a common-sense intervention. Based on some preliminary analyses of a research project I am conducting at the moment, it does appear that increased recreational reading improves reading. This would be the first large-scale demonstration that recreational reading practice does improve reading ability.

Based on this evidence (albeit admittedly tentative) I would recommend that comprehension interventions include opportunities to do independent, non-instructional reading, as well as practice in reading material that has instructional relevance.

Another dimension of practice is oral reading fluency. While there is evidence for younger readers that fluency is related to comprehension, it also appears that this relationship may be attenuated for older students. If this is the case, interventions should include fluency training for younger readers, with decreasing emphasis for students in upper grades. However, the assessment system should also sort readers into appropriate instruction depending on fluency needs.

## Professional Development

Any intervention will require some learning on the part of teachers if they are to deliver the intervention effectively. However, as noted earlier, the requirements of the program should not be so high as to make the implementations prohibitive either in time or other resources.

Professional development has to be ongoing, intensive, and tailored to local contexts. Interventions can be derailed without appropriate professional development. With appropriate professional development, interventions can be made even more successful. Intervention fidelity will be difficulty to maintain without professional development.

## Developmental Nature of Content Reading

Content reading instruction requires attention to the notion that readers continue to develop even after the reach middle and high school. The work of Alexander and her colleagues show that at least some reading in content materials is

Alexander and Jetton (2000) proposed a developmental view of learning from text. They emphasized the notion that the ability to learn from text changes over the course of one's education (and, presumably, over the remainder of one's life experiences).

In this perspective, readers progress from an *acclimated* stage, in which the emphasis is on orientation and adaptation. Learners are attempting to understand the structure of an unfamiliar domain of information. An important characteristic of this stage is that students often apply strategies inefficiently because they have limited subject matter knowledge.

At the next, more advanced stage, learners are termed *competent*. To reach this stage (and not all learners do), readers have to develop a sufficient level of knowledge, strategic capability, and motivational interest and goals. This competence entails both quantitative and qualitative transformations in knowledge of the domain and strategic processing. As students reach this level, their deeper level of subject matter knowledge facilitates the acquisition of new knowledge. Important among the factors that account for the transition to competence from acclimation is the use of strategies (Alexander and Murphy, 1999). However, in the end, motivation was the clearest determiner of successful students.

The highest stage in this perspective is labeled *proficiency* or *expertise*. At this stage, readers have a great deal of knowledge of specific domains, deep interest in the topic, and a desire to explore or learn more about the domain. Alexander and Jetton suggest that few students ever reach this final stage.

On the basis of the research that supports this developmental perspective, they offer instructional implications that include the need for differential instructional support for learning, depending on the stages students have attained. Quality of text is differentially related to students' abilities to learn. Students at the acclimated stage are most affected by poor texts, while students who have reached more advanced stages can compensate for flaws in the textbooks. Finally, Alexander and Jetton suggest that instruction encourages learner autonomy and intrinsic motivation.

## ELL Populations

There are many variables that haven't been included in the material above. Foremost among them are the issues of interventions for ELL populations. Instructional accommodations need to be made for this population, although the exact accommodations are not well established, since many of the recommendations depend on the nature of the native language of the learner. That is, the recommendations would differ depending on whether the first language was alphabetic with a Roman orthography or not. Languages that contain many cognates with English (e.g. Spanish) will require different accommodations from languages without cognates (e.g. Hmong).

The problem is less acute for students who are literate in a first language, since there is a great deal of transfer from first language literacy to second language literacy. For younger students who come to the task without a first literacy, the interventions should probably focus on oral language development in both first and second language. Writing should be stressed for older students, since there is some evidence that levels of writing for ELLs may surpass oral proficiency estimates. It still seems to be an open question whether first literacies should be taught in first languages, when the target skill is second language literacy, depending on the transfer effect to make learning more effective. Regardless, interventions have to account for these variations.

## Concluding Comments

The above ruminations often seem to suggest that there is, or might be, a "perfect" intervention for comprehension. It is unlikely that we will find such an intervention, at least not in the near future. All of the characteristics mentioned above are desirable. Many of them do not have to be part of instruction for all students, stressing the crucial importance of a comprehensive assessment system to guide the implementation any intervention. Comprehension interventions are not full-blown reading programs, even though they must address all the reading skills of comprehensive programs. Most of the time, comprehension interventions will have to do assessment so that impediments to comprehension can be remediated. In general, it is likely case is that comprehension interventions will be modular and will be targeted as suggested in the discussion above.

There should be some effort to incorporate the latest developments in computer technology to deliver the intervention. With the great advances in speech recognition and computerize analysis of written responses, computer technology is showing ever greater promise. Computerized delivery offers the added advantage of being able to target individual students with the appropriate kinds and amounts of practice and leverage teacher time by allowing individuals to continue instruction and practice when the teacher is occupied with other students. Further, the multimedia capabilities make radically different ways of literacy instruction possible.

Finally, the evaluation of these types of interventions should be done with something like a systems approach. That is, the intervention should be viewed as a component of a larger educational context. The introduction and use of any intervention will alter the

other components of the context and the interactions among them. Consequently, there should be stricter requirements for evaluations of interventions than simply being tested against a control group. For example, if a comprehension intervention produces far better readers, the remainder of the system will have to accommodate those better readers (with different instructional strategies in non-intervention settings) if their gains are to be maintained.

REFERENCES

Alexander, P. A., & Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson & R. Barr (Eds.), *Handbook of Reading Research* (Vol. III, pp. 285-310). Mahwah, NJ: Lawrence Erlbaum Associates.

Alexander, P. A., & Murphy, P. K. (1999). Learner profiles: Valuing individual differences within classroom communities. In P. L. Ackerman & P. C. Kyllonen (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 413-436). Washington, DC: American Psychological Association.

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. *Advances in Reading/Language Research, 2*, 231-256.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. Reading Research Quarterly, 23(3), 285-303.

Curtis, M. (In press.) Adolescents who struggle with word identification: Research and practice. To appear in T. Jetton and J. Dole (eds.), *Adolescent Literacy Research and Practice*. New York: Guilford.

Guthrie, J. T. & Wigfield, A. (2000). Engagement and Motivation in Reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R Barr (Eds.) *Handbook of reading research, Vol. III.* (pp. 403-422). Mahwah, NJ: Lawrence Erlbaum Associates.

Hart, B. & Risley, T. (1995). *Meaningful Differences in Everyday Parenting and Intellectual Development in Young American Children*. Baltimore: Brookes.

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

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

Sticht, T. G., Beck, L. B., Hauke, R. N., Kleiman, G. M., & James, J. H. (1974). *Auding and reading: A developmental model*. Alexandria, VA: Human Resources Research Organization.