Methodology

THIS MATERIAL IS EXTRACTED FROM *CSRQ CENTER REPORT ON ELEMENTARY SCHOOL COMPREHENSIVE SCHOOL REFORM MODELS*

http://www.csrq.org

NOVEMBER 2005

AMERICAN INSTITUTES FOR RESEARCH®

The contents of this report were developed by the Comprehensive School Reform Quality (CSRQ) Center, which is operated by the American Institutes for Research (AIR), under a grant from the U.S. Department of Education. The views expressed in this report do not necessarily reflect those of the U.S. Department of Education, AIR, or the CSRQ Center.

Methodology

Although this report is intended for a general readership, cutting edge scientific concepts and processes have been applied to produce the reviews. In this section, we detail the research methods used to support these reviews. This section highlights some of the challenges posed in conducting systematic reviews of evidence and gives our technical readers the background needed to judge the quality of our scientific efforts.

Past systematic reviews of model effectiveness in comprehensive school reform (CSR) have relied heavily on unpublished or published reports on specific CSR models—most notably the work by Borman, Hewes, Overman, & Brown (2002) and Herman et al. (1999), which compared the effectiveness of specific CSR models in raising student achievement. The Comprehensive School Reform Quality (CSRQ) Center's work builds on this work to quantitatively evaluate CSR models as well as to provide qualitatively a narrative description of each reviewed model.

The CSRQ Center's researchers recognize that, while student achievement is critical, education consumers also rely on thorough descriptions of CSR models and want to know how their school may change if they implement a specific model. School staff also seeks information about the experience of other schools implementing CSR models. The CSRQ Center's approach combines qualitative and quantitative research techniques to report on CSR models' impact on student achievement and on experiences of schools implementing these models. Creswell (1994, p. 175) advocated the use of multimethods by stating five purposes:

1. Triangulation, in seeking convergence of results
2. Complementary, in that overlapping and different facets of a phenomenon may emerge
3. Developmentally, wherein the first method is used sequentially to help inform the second method
4. Initiation, wherein contradictions and fresh perspective emerge
5. Expansion, wherein the mixed methods add scope and breadth to a study

The CSRQ Center strives to replicate past analyses by determining student achievement effects and to expand and fully describe each component of a CSR model and the services it offers to schools.

As described in the introduction, the CSRQ Center developed the Quality Review Tool (QRT), a three-part, multimethod tool to collect and analyze qualitative and quantitative data to evaluate CSR models for the education consumer.

1. **QRT Part 1** is the *qualitative data collection* phase. The purpose of QRT Part 1 is to gather (a) supporting information from the CSR model's directors and three school principals and (b) descriptive information about the CSR model, such as professional development, technical assistance, and research-based design.
2. **QRT Part 2** is the *quantitative data collection* phase. The purpose of QRT Part 2 is to conduct a systematic review of the literature on the effectiveness of a CSR model on student achievement, other outcomes, such as attendance and graduation rates and family and community involvement outcomes.
3. **QRT Part 3** is the *data analysis* phase, in which the qualitative and quantitative data are synthesized to generate effectiveness ratings of the CSR model. These ratings (Very Strong, Moderately Strong, Moderate, Limited, Zero, and No Rating) are

developed for several categories including evidence of positive effects on student achievement, additional outcomes, and parent, family, and community outcomes; evidence of a link between research and the model's design; and evidence of the model's ability to provide services and support (e.g., readiness and professional development/technical assistance) to schools to enable successful implementation.

Sample of Elementary School CSR Models

The CSRQ Center gathered a list of more than 100 elementary school CSR models by consulting previous reviews (Borman et al., 2002; Herman et al., 1999; Slavin & Fashola, 1998), the Southwest Educational Development Laboratory's (SEDL) *CSR Awards Database*, and the Northwest Regional Education Lab's (NWREL) *Catalog of School Reform Models Database*. From this list, we selected a final sample by

1. Determining market share, as defined by the total number of schools implementing the CSR model;
2. Exploring the replicability of the CSR model, as determined by geographic spread; and
3. Investigating the comprehensiveness of the CSR model's design.

Each step of the information gathering process consulted previous reviews, databases, and the Web sites of the CSR model providers.

For Step 1 (market share), CSRQ Center's researchers searched the CSR model provider's Web site for information on the total number of schools that used the CSR model. This information was verified using the SEDL's CSR Awards database. From the list of more than 100 CSR model providers, the number of schools using a particular CSR model ranged from 1 school to several hundred schools. The selection criterion for market share was to include CSR models that were used in 20 or more schools. This yielded 54 CSR model providers.

For Step 2, (replicability), CSRQ Center's researchers consulted information from the CSR model provider's Web site and the SEDL's CSR Awards database to determine whether the 54 CSR models from Step 1 were present in three or more states. This step narrowed down the list from 54 to 49 CSR model providers.

For Step 3 (comprehensiveness), CSRQ Center's researchers examined whether the CSR model's design features met components identified by the U.S. Department of Education: governance, technical assistance, classroom practices, professional development, leadership development, benchmarks/assessments, and curriculum (U.S. Department of Education, n.d.). For coding purposes, components were defined as follows:

- **Governance** was defined as operations and management conducted in schools. Key words associated with governance were operations, structure, management, scheduling, committees, blocks, and administration.
- **Technical assistance (TA)** was defined as classroom operational or management assistance through mentoring, coaching, or other services provided to teachers. Key words associated with TA were troubleshooting, coaching, and mentoring.
- **Classroom practices (CP)** was defined as pedagogical, structural, and behavioral management practices that a teacher uses in a classroom. Key words associated with CP were pedagogy, classroom management, classroom structure, teaching strategies, and philosophy of instruction.
- **Professional development (PD)** was defined as teacher training on a specific topic. This training typically occurs in a workshop or conference environment. Key words associated with PD were training (on specific topics), conferences, and workshops.

- **Leadership development (LD)** was defined as administrative training or development for school personnel in leadership positions (e.g., principals, grade-level chairs, and lead teachers). Key words associated with LD were leadership training and/or development.
- **Benchmarks/assessments** was defined as tests and evaluations used to measure students' skills and understanding and academic progress. Key words associated with benchmarks/assessment were measurable goals, formative evaluation, and benchmarks of progress.
- **Curriculum** was defined as the scope and sequence of learning objectives and indicators, as well as material provided for lessons to instruct such objectives. Key words associated with curriculum were materials, scope and sequence, standards, and learning objectives.

Each CSR model was given a point for each component or criterion the model met based on information found on the model's Web site and additional resources including but not limited to *An Educator's Guide to Schoolwide Reform* (Herman et al., 1999), *Show Me the Evidence* (Slavin & Fashola, 1998), and the following Web sites: http://www.ed.gov, http://www.SEDL.org, and http://www.nwrel.org. Each CSR model provider that had five or more components in its design was included in the final sample. This step narrowed the list from 49 to 22 CSR models for review.

QRT Part 1: Qualitative Data Collection Phase

QRT Part 1 is the qualitative data collection phase. It includes guidelines for conversations with model directors and school principals and the collection of artifacts from CSR models and schools and additional information about the CSR model from publicly available resources (Bogdan & Biklen, 1998; Creswell, 1994, 1998).

QRT Part 1, including the guidelines for phone conversations, conversation questions, and artifact lists, was pilot tested with one of the CSR model providers in the sample. Based on feedback from the pilot conversations, researchers at the CSRQ Center modified the qualitative data collection process. An experienced and trained qualitative researcher at the American Institutes for Research (AIR) provided training on information gathering techniques, coding artifacts, and synthesizing qualitative data to develop a complete description of each CSR model in the sample. The qualitative researchers met weekly to ensure consistency across the qualitative data collection efforts.

For QRT Part 1 (qualitative data collection), qualitative researchers performed four main steps:

1. **Complete an initial description of the CSR model description by using a standardized form.** The CSRQ Center developed the Model Description Form, a comprehensive survey instrument for compiling existing information about a CSR model, including mission, history, market share, costs to the school, and design of each of the CSR components as outlined by the U.S. Department of Education. For example, researchers gathered information about the CSR model's organization and governance, such as how the CSR model provides site-based autonomy, whether additional personnel are needed, and whether the CSR model requires changes to the structure of the school. For questions about professional development, researchers gathered information about which school personnel are required to attend professional development; what types of professional development are offered prior to, during, and after implementation; and what strategies are available to help a school build capacity to provide its own professional development. In all, researchers gathered information about the CSR model's organization and governance, professional development, technical assistance, curriculum, instruction, inclusion, technology, time and scheduling, instructional

grouping, student assessment, data-based decision making, and parent, family, and community involvement. The researchers also requested benchmarks and explicit citations that link the model's design to a research base. The researchers completed this survey using the CSR model provider's Web site and other publicly available information.

2. **Conduct a phone conversation with the provider of the CSR model to verify previously gathered information.** Conversations were structured around the Model Description Form (completed in step 1). On average, phone conversations lasted 90 minutes.

3. **Conduct phone conversations with three school principals who use the CSR model.** The conversations verified information gathered in steps 1 and 2. Schools were randomly selected from a list provided by the CSR model's provider or through the SEDL CSR Award database. The conversations were guided by the Model Description Form.

4. **Complete a final description of the CSR model by using a standardized form.** The Model Description Form-Complete synthesized all sources of qualitative data gathered, such as the conversations with the model's provider and the three school principals and artifacts collected from the CSR model provider or schools. The Model Description Form-Complete was checked for quality control twice to ensure that each item had 100% agreement between the two qualitative researchers. This form was then used to organize the data through the identification of core components. Core components are considered essential to the successful implementation of the model according to the CSRQ Center's standards. Additionally, these data were coded to answer several questions:

   - Is there a strong link between research and the CSR model?
   - Does the CSR model track and support full implementation in all schools?
   - Does the CSR model help schools allocate resources to implement the model?
   - Does the CSR model provide comprehensive training opportunities and supporting materials?
   - Does the CSR model develop the schools' internal capacity to provide professional development?

QRT Part 2: Quantitative Data Collection Phase

QRT Part 2 is the quantitative data collection phase. Using systematic review methods (Borman et al., 2002; Lipsey & Wilson, 2001), QRT Part 2 includes protocols to conduct systematic literature reviews and to code research studies for statistical and causal validity information.

QRT Part 2, including the protocols for literature reviews and coding instruments, was pilot tested using the same CSR model provider from the qualitative data collection efforts (QRT Part 1). Based on feedback from the pilot test, the process for conducting the literature review was improved and the coding instruments were refined. An experienced and trained quantitative researcher at AIR conducted training on how to use the coding instruments to ensure consistency in the data collection. The training included a presentation of the definitions of different research designs, causal validity issues, and background information on effect size calculations.

For QRT Part 2, quantitative researchers completed five main steps:

1. **Conduct a thorough literature search.** For each CSR model, quantitative researchers searched educational databases (e.g., JSTOR, ERIC, EBSCO, Psychinfo, Sociofile, NWREL, DAI), Web-based repositories (e.g., Google, Yahoo, Google Scholar),

and two previous studies on comprehensive school reform (Herman et al., 1999; Borman et al., 2002). From these sources, quantitative researchers screened for *initial relevance* nearly 800 article abstracts or summaries across the 22 models in the sample. To pass the initial screen, the sources had to meet several criteria: be published or distributed between 1980 and April 2005, examine at least one of the CSR models being investigated, use quantitative methods, and be reported as a full-text research paper (i.e., not a PowerPoint presentation or executive summary). From these articles, researchers identified 407 studies to code. Of those, 360 were retrievable and available for coding. Appendix X provides a summary table of the number of studies that passed through each phase of the QRT Part 2 process.

2. **Complete a Study Description Outcome Form (SDOF), the first standardized coding sheet.** The CSRQ Center's quantitative researchers used the SDOF to code and document each source's research design, outcome variables, and demographic information. The Center assigned a lead and secondary coder for each source. The SDOF was completed by the lead coder. Then, the secondary coder verified all the information for 100% agreement. At this stage of coding, the primary focus was to screen each source for a reliable research design. Studies that *were not eligible for full review* were often evaluations of implementation theories supporting the CSR model with no quantitative data on outcomes or used research designs that were not sufficiently rigorous (e.g., one group pretest-posttest research designs). Research designs that passed this stage included experimental designs and quasi-experimental research designs with both pre- and posttests that evaluated the CSR model with a control group (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) and longitudinal and cohort designs with multiple testing periods. Studies with research designs that passed this screen and included student achievement outcomes became eligible for full review. A total of 117 studies passed this step and were eligible for full coding in step 3.

3. **Complete the Quality Indicators Form (QLIF), the second standardized coding sheet.** Researchers used the QLIF to code studies that appeared to use rigorous research designs. The QLIF served two purposes: It examined the quality of the research and gathered statistical information. Researchers examined the quality of the research, such as the internal and external validity, face and psychometric validity of the outcome measures, and other quality indicators (Herman et al., 1999). Coders also collected statistical information, such as effect sizes reported by the authors or raw statistical information. For each study that was relevant for full review, two quantitative researchers independently coded one QLIF for each achievement outcome in a study.

4. **Reconcile the two QLIF coding sheets to attain 100% agreement on each coded item.** If the two quantitative researchers could not reach a consensus, a review coordinator reviewed the coding sheets to facilitate reconciliation. After the reconciliation process, a final QLIF reflected the 100% agreement.

5. **Rate each article on an overall causal validity score.** The final step was to systematically map the information from the final QLIF (the reconciled version) based on a set of rubrics designed to score each study for its causal validity (Shadish et al., 2002) as *inconclusive*, *suggestive*, or *conclusive*. Studies determined to be suggestive or conclusive met CSRQ Center standards for rigor of research design.

   A study was inconclusive if it had critical threats to validity, such as using testing instruments with poor face validity and reliability, insufficient program fidelity, nonequivalence of treatment/control groups, lack of proper baseline, and/or timing of outcome measures (less than 1 school year after CSR model implementation or less than 1 academic

year elapsed between pretest and posttest). Noncritical threats to validity include historical events, disruption/novelty effects, instrumentation changes, maturation, selection bias, and statistical regression (Shadish et al., 2002).

Suggestive studies had zero critical threats but more than two noncritical threats. Studies without control groups including longitudinal and cohort research designs were capped at suggestive, unless the analytic techniques generated higher levels of rigor.[1] Conclusive articles had higher levels of rigor, that is, experimental and quasi-experimental designs that had zero critical threats to validity and fewer than two noncritical threats to validity. Effect sizes were reported or calculated only from studies that had a conclusive causal validity rating (Cooper, 1998; Light & Pillemer, 1984; Shadish et al., 2002). If the researcher could not calculate an effect size because of missing data, then the researcher conducted one of the following steps: (a) contacted the author for the statistical information needed, (b) imputed missing data, particularly standard deviations and sample size using protocols established in previous meta-analysis (Borman et al., 2002), or (c) chose not to include the study in the synthesis if options a and b were not feasible.

QRT Part 3: Data Analysis Phase

QRT Part 3 synthesizes the qualitative and quantitative data to evaluate each CSR model in five main categories.

1. Evidence of positive effects on student achievement:
    a. Evidence of positive overall effects
    b. Evidence of positive effects for diverse student populations
    c. Evidence of positive effects for specific subject areas
2. Evidence of positive effects on additional outcomes (e.g., student discipline, student attendance, school climate, retention/promotion rates, and teacher satisfaction)
3. Evidence of positive effects on parent, family, and community involvement
4. Evidence of a link between research and the model's design
5. Evidence of services and supports to schools to enable successful implementation:
    a. Evidence of readiness for successful implementation
    b. Evidence of professional development/technical assistance for successful implementation

Category 1 uses the quantitative information gathered in QRT Part 2. For each CSR model in the sample, the quantitative information—including the number of studies coded, the number of studies that were rated as suggestive and conclusive, the percentage of findings in the suggestive and conclusive sources that demonstrated a positive impact, and the average effect size of those significant findings—was mapped onto rubrics to determine if the model should receive a very strong, moderately strong, moderate, limited, zero, or no rating for effects on student achievement. Quantitative researchers systematically aggregated results according to the QRT 3 rubric for the overall effect by grade, subject (reading, writing, math, science, and social studies), and diverse student populations (e.g., high poverty, minority, learning disabled and other special needs, and urban and rural students).

[1] For example, backward-looking interrupted time series designs were considered more rigorous than longitudinal or longitudinal cohort studies that examined trends over time.

Category 2 evaluates the positive effects of each CSR model on additional outcomes, and Category 3 evaluates the evidence of positive effects of each CSR model on parent, family, and community involvement. Similar to Category 1, quantitative researchers mapped onto rubrics the information about the number of sources (that evaluated these outcome variables), the number of sources that were suggestive and conclusive, the percentage of findings that demonstrated a positive impact, and the average effect size of those positive findings.

In general, the rubrics for the quantitative information for Categories 1–3 are as follows:

- **Very Strong.** If a model had at least 10 studies that met CSRQ Center's standards for rigor of research design with at least 5 rated conclusive (and/or conclusive studies constitute at least 50% of the total studies coded) and 75% of the outcomes showed statistically significant positive model effects ($p \leq .05$), with an overall mean model achievement effect of at least ES = +0.25, then the model received a very strong rating, which is symbolized by a fully shaded circle (●).
- **Moderately Strong.** If a model had 5 to 9 studies that met CSRQ Center's standards for rigor of research design with at least 3 rated conclusive (and/or conclusive studies constituted at least 50% of the total studies coded) and 51% to 74% of the outcomes showed statistically significant positive model effects ($p \leq .05$), with an overall mean program achievement effect of ES = +0.20 to +0.24, then the model received a moderately strong rating, which is symbolized by a three-fourths shaded circle (◕).
- **Moderate.** If a model had 2 to 4 studies that met CSRQ Center's standards for rigor of research design with at least 1 rated conclusive (and/or conclusive studies constituted at least 50% of the total studies coded) and 26% to 50% of the outcomes showed statistically significant positive model effects ($p \leq .05$), with an overall mean model achievement effect of ES = +0.15 to +0.19, then the model received a moderate rating, which is symbolized by a half-shaded circle (◑).
- **Limited.** If a model had 1 study that met CSRQ Center's standards for rigor of research design and 1% to 25% of the outcomes showed positive model effects that were statistically significant ($p \leq .05$), then the model received a limited rating, which is symbolized by a one-fourth shaded circle (◔).
- **Zero.** If a model had zero studies that met CSRQ Center's standards for rigor of research design or 0% of the outcomes in the studies that met CSRQ Center's standards for rigor of research design showed statistically significant positive effects, as required for a limited rating, then the model received a zero rating, which is symbolized by a circle with a horizontal slash (⊘).
- **Negative.** If a model had at least 10 studies that met CSRQ Center's standards for rigor of research design with at least 5 rated conclusive (and/or conclusive studies constituted at least 50% of the total studies coded) and 75% of the outcomes showed statistically significant negative model effects ($p \leq .05$), with an overall mean model achievement effect of ES < 0, then it received a negative rating, which is symbolized by a circle with a minus sign (⊖). This indicated that research suggests the model has detrimental effects. In practice, this review did not find any evidence of this kind for any model.
- **No Rating.** If a model had no studies (i.e., no evidence was available), then the model received a no rating, which is symbolized by a circle with "NR" (Ⓝⓡ).

Category 4 evaluates the link between research and the CSR model's design. This category uses the qualitative information from QRT Part 1. Qualitative researchers applied the information synthesized in the Model Description Form (from QRT Part 1) into the following rubric.

- **Very Strong.** If a model provided documentation that explicitly described and convincingly supported links between the research base and *all* (100%) core components of its design, then it received a very strong rating, which is symbolized by a fully shaded circle (●).
- **Moderately Strong.** If a model provided documentation that explicitly described and supported links between the research base and *most* (75%) of the core components of its design, then it received a moderately strong rating, which is symbolized by a three-fourths shaded circle (◕).
- **Moderate.** If a model provided documentation that explicitly described and supported links between the research base and *half* (50%) of the core components of its design, then it received a moderate rating, which is symbolized by a half-shaded circle (◑).
- **Limited.** If a model provided documentation that explicitly described and supported links between the research base and *less than half* (below 50%) of the core components of its design, then it received a limited rating, which is symbolized by a one-fourth shaded circle (◔).
- **Zero.** If a model provided documentation that referred to a *nonspecific* research base to support the inclusion of the core components in its design, then it received a zero rating, which is symbolized by a circle with a horizontal slash (⊘).
- **No Rating.** If the CSRQ Center was unable to conduct a conversation with the model provider or obtain complete information to verify evidence, then the model received a no rating, which is symbolized by a circle with "NR" (Ⓝⓡ).

Two main questions guided the ratings for Category 5 (evidence that the model provider offers services and support to schools to ensure successful implementation). The first question—does the CSR model provide evidence of readiness for successful implementation—included the following subcategories:

- Provider ensures initial commitment from schools.
- Provider tracks and supports full implementation in schools.
- Provider helps schools allocate resources needed to fully implement the CSR model.

Qualitative researchers used the information synthesized in the Model Description Form (from QRT Part 1) to rate the three subcategories using a specific rubric. Next, these three ratings were averaged to determine the rating for evidence of readiness for successful implementation. In general, a model's rating was based on evidence of the following: a formal or informal process for establishing an initial understanding of the model, strategies to develop faculty buy-in, formal or informal benchmarks for all or some of its core components, and a formal or informal process for the allocation of such school resources as materials, staffing, and time.

The second question—does the CSR model provide schools with professional development and technical assistance needed to help teachers implement the model—included the following subcategories:

- Provider offers comprehensive training opportunities and supporting materials.
- Provider ensures that professional development effectively supports full model implementation.
- Provider develops school's internal capacity to provide professional development.

Again, each subcategory received a rating. The three ratings were averaged to determine the rating for evidence of professional development and technical assistance for successful implementation. In general, a model's rating was based on evidence of the following: a variety of training opportunities, supporting materials for professional development in all or some of its core components, and a formal or informal plan to help build a school's capacity to provide professional development.

In addition to the ratings across these five categories, the qualitative data gathered in QRT Part 1, such as the artifacts and phone conversations, were synthesized into a narrative description of each CSR model. Each narrative includes in-depth information about the CSR model's costs and descriptions of the following components: organization and governance; curriculum and instruction; scheduling and grouping; technology; monitoring of student progress; parent, family, and community involvement; professional development and technical assistance; and implementation expectations and benchmarks.

In all, qualitative and quantitative data were mapped to rate a CSR model on

- Evidence of positive effects on student achievement;
- Evidence of positive effects on additional outcomes;
- Evidence of positive effects on parent, family, and community outcomes;
- Evidence of link between research and the model's design; and
- Evidence of services and support to schools to enable successful implementation.

The quantitative data provided a systematic literature review of the reported effects of student achievement and other outcome variables. CSR models that have relatively more literature consisting of evaluation studies were more likely to achieve higher ratings in Categories 1–3 (as long as results demonstrated positive impact). Furthermore, by using qualitative data, newer CSR models or those that do not have a substantial number of evaluation reports can be evaluated on dimensions such as professional development. Although past research on student achievement offers important considerations for education consumers, they may also consider whether the CSR model's design is based on solid research and provides a strong commitment to support schools through professional development and technical assistance. Newer models may not have had sufficient time to conduct research on their effectiveness, but they ought to be able to clearly demonstrate that they *can work*, that is, that the model's design is based on solid evidence of *what works*. Hence, by using both qualitative and quantitative methods, the CSRQ Center strives to provide the education consumer with a thorough and systematic description of the effectiveness of each CSR model reviewed in this report.

By using qualitative and quantitative methods to evaluate the effectiveness of widely implemented CSR models, this study also strives to provide usable information to education consumers. U.S. Education Secretary Margaret Spellings recently stated that the No Child Left Behind Act "rests on the common sense principles of accountability for results, data-based decision making, high expectations for all, and empowering change" (U.S. Department of Education, 2005).

Meeting these goals will require a significant expansion of information for education consumers about what works. This report is intended to act as a decision-support tool for educators wishing to find effective CSR approaches for meeting locally defined needs. It helps to provide such information and will help increase its use in education decision making—marking a significant change in the culture of the education system to meet the needs of educators, policymakers,

community leaders, families, and most importantly, America's children.

References

Bogdan, R. C., & Biklen, S. K. (1998). *Qualitative Research for Education: an introduction to theory and methods* (3rd ed.). Boston: Allyn and Bacon.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2002). *Comprehensive school reform and student achievement: A meta-analysis.* Baltimore, MD: Center for Research on the Education of Students Placed At Risk.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cooper, H. (1998). *Synthesizing research* (3rd. ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches.* Thousand Oaks, CA: Sage.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions.* Thousand Oaks, CA: Sage.

Herman, R., Aladjem, D., McMahon, P., Masem, E., Mulligan, I., O'Malley, A. S., et al. (1999). *An educators' guide to schoolwide reform.* Arlington, VA: Educational Research Service.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton Mifflin.

Slavin, R. F., & Fashola, O. S. (1998). *Show me the evidence! Proven and promising programs for America's schools.* Thousand Oaks, CA: Corwin Press.

U.S. Department of Education. (n.d.). *Comprehensive School Reform Program*. Retrieved November 8, 2005, from http://www.ed.gov/programs/compreform/2pager.html

U.S. Department of Education. (2005, March 10). *Expanding the promise, continuing the progress.* Retrieved November 2, 2005, from http://www.ed.gov/news/speeches/2005/03/03102005.html