Teaching Writing to At-Risk Students: The Quality of Evidence for Self-Regulated Strategy Development

Scott K. Baker
Pacific Institutes for Research/University of Oregon

David J. Chard
Southern Methodist University

Leanne R. Ketterlin-Geller
Chanisa Apichatabutra
Christian Doabler
University of Oregon

Abstract: This study evaluates the quality of the research and evidence base for a writing intervention called Self-Regulated Strategy Development (SRSD; Graham & Harris, 1989; Harris & Graham, 1996) for students with and at risk for learning disabilities, using criteria for group research studies suggested by Gersten et al. (2005) and single-subject research studies suggested by Horner et al. (2005). Five experimental and quasi-experimental studies and 16 single-subject studies investigating SRSD were analyzed on numerous methodological dimensions. Both the group design and single-subject studies also met proposed standards for an evidence-based practice. The potential value of analyzing approaches and interventions using the proposed quality indicators and standards for evidence-based practices is discussed, as are implications for research and practice.
students with learning disabilities (LD; Englert, Raphael, Anderson, Anthony, & Stevens, 1991; Gersten & Baker, 2001; Graham & Harris, 1997).

Few educators question the value of directly teaching students to write effectively. Yet factors such as the amount of time students spend being taught systematically how to write seem to conflict with the importance educators attach to writing (Graham & Harris, 1997). For example, writing instruction receives much less instructional focus than does reading or mathematics (Baker, Gersten, & Graham 2003). Fragments of writing instruction may be incorporated within reading or content-area instruction, but sustained and cohesive writing instruction is not particularly common in school settings (Graham & Harris, 1997). Further encroachments on time devoted specifically to writing instruction may occur as schools increasingly search for ways to allocate additional time for reading instruction.

In the past, it was common for educators to think of writing instruction somewhat passively, consisting mainly of having students read extensively and encouraging them to apply to their own writing what they observed in the writing of others. Research on these types of exposure methods indicates that they do not help students become better writers, leading to an era of advocacy for more explicit approaches (Hillocks, 1984).

**CONSEQUENCES OF POOR WRITING**

Although the importance of fostering effective writing skills among students is unquestioned, there is clear evidence from the National Assessment of Educational Progress (NAEP) that these efforts are insufficient (Graham & Perin, 2007). On the NAEP writing assessment for 2002, students in Grades 4, 8, and 12 wrote narrative, informative, and persuasive essays, and their performance was categorized as Below Basic, Basic, Proficient, or Advanced. Basic is defined as "partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade" (Institute of Education Sciences, 2004). In other words, students who score at Below Basic and Basic levels are not meeting minimum standards for competent writing. On the 2002 NAEP, in Grades 4, 8, and 12, 72%, 69%, and 77% of students respectively wrote at Below Basic and Basic levels (Graham & Perin, 2007).

On the NAEP 2007 report, which provides writing results for Grades 8 and 12 (Institute of Education Sciences, 2007), there were slight increases overall in the percentage of students in the proficient category and above, and for each demographic subgroup. However, for students with disabilities, the outcomes are troubling. Ninety-four percent of students with disabilities scored in the Basic and Below Basic categories. In other words, only 6% of students with disabilities were considered to have proficient writing skills. In summarizing research on writing instruction for students with LD, Gersten and Baker (2001) stated that on "every conceivable measure of writing performance—including both measures of writing quality and quantity and occurring across narrative and a range of expository text structures—students with learning disabilities write much more poorly than do students without disabilities" (p. 252). This finding has been consistently supported (Englert et al. 1991; Graham, 1990; Graham & Harris, 1997; Wong, Butler, Ficzere, & Kuperis, 1996, 1997).

The value of writing effectively is not confined to accomplishing academic tasks in school settings. Recent reports by the National Commission on Writing (2004, 2005) indicated that the majority of public and private employers state that writing proficiency is critical in the workplace and directly influences their hiring and promotion decisions. Writing is no longer a requirement limited to the daily tasks of professionals, but it is a workplace demand that extends to nearly all living-wage jobs. The lack of fundamental writing skills among new employees is such that the National Commission on Writing estimated that 30% of employers require on-the-job training in basic writing skills. The financial cost of this workplace remediation is considerable. Private companies spend approximately $3.1 billion annually on writing remediation, and state governments spend about $221 million annually (Graham & Perin, 2007; National Commission on Writing, 2005).
The consequences of illiteracy, including problems students experience with fundamental writing skills, have spurred research on writing instruction in K to 12 settings. Although this research base is not as extensive as the research on reading instruction, many studies have been conducted and special education researchers have played leading roles in these efforts (see Baker et al., 2003 and Gersten & Baker, 2001, for summaries of this research).

The research on writing instruction that has blossomed in the last 20 years has increasingly focused on the quality of writing content rather than writing mechanics (Cersten, Baker, Pugach, Scanlon, & Chard, 2001). A catalyst for this increase of studies on writing quality was the meta-analysis on writing instruction by Hillocks (1984). Hillocks concluded that effective writing instruction had clear and specific objectives and prepared students to write about specific topics. Writing instruction that included planned brainstorming activities and that helped students organize information prior to writing was more effective than methods that ignored or gave short shrift to writing preparation. More traditional writing instruction methods, such as combining simple sentences into more complex sentences, were considerably less effective than methods that addressed the full range of the writing process. Particularly ineffective were methods that had students write substantial amounts of text with minimal guidance from the teacher, or methods that had students attempt to emulate features of good writing that they found in the writing of others. The least effective approaches focused on studying parts of speech and sentence fragments.

Since Hillocks' (1984) meta-analysis, researchers have increasingly focused on ways to improve writing content and have embedded methods within writing instruction across multiple stages of the writing process. Different researchers have identified different numbers of stages, but essentially the stages are: (a) instruction in planning to write (e.g., Englert et al., 1991); (b) writing from well-developed plans of action (Gersten & Baker, 2001; Gersten et al., 2001); and (c) revising initial written drafts (e.g., MacArthur, Schwartz, & Graham, 1991).

Recently, Graham and Perin (2007) conducted a meta-analysis on writing interventions for students in Grades 4 to 12 and found 142 studies that met their inclusion criteria. From these studies, they calculated 176 effect sizes. This magnitude of research is in sharp contrast to the 29 effect-size calculations Hillocks (1984) derived for similar types of studies targeting students in Grades 4 to 12. Based on their findings, Graham and Perin described 11 elements of effective instruction in Grades 4 to 12. Although the elements were separated in the analysis, many of the elements overlapped and included multiple stages of the writing process. For example, in studies on collaborative writing approaches, students are taught to work together to plan, draft, and revise their writing samples. Studies on teaching students writing strategies also focus on these stages of writing. Other studies honed in on specific stages of writing, such as approaches that teach sentence combining, where students are taught how to write more complex sentences. Graham and Perin encouraged readers not to consider the elements "as isolated but rather as interlinked" (p. 11). It is the linkage of the elements that leads to comprehensiveness in writing instruction.

Less research has been conducted on students with LD specifically. Examining primarily published studies (Graham and Perin, 2007, reviewed a much broader range of studies), Gersten and Baker (2001) conducted a meta-analysis involving 13 experimental and quasi-experimental studies of interventions designed to improve writing content in a variety of genres with students with LD. The analysis focused specifically on how the writing process was taught, and the results indicated that teaching writing strategies to students with LD could result in considerable improvements in writing quality. Although single-subject studies were not targeted in this meta-analysis formally, the informal examination of single-subject studies also supported this conclusion about the positive benefit of teaching writing strategies directly to students with LD.

A comprehensive approach to writing instruction that has been used with students with and without disabilities has been developed by Graham and Harris and their colleagues (Harris...
Over the past 25 years, Graham and Harris have developed an approach to writing instruction that focuses in particular on the role of self-regulation in the development of written compositions. Although the approach can be used with all students, many of the features are particularly aligned with the needs of students with LD or other students who struggle with writing tasks. In their approach, called Self-Regulated Strategy Development (SRSD), students learn specific strategies for planning, drafting, and revising text. Explicit and strategy-based instruction is fundamental to this approach. Instruction occurs across the following six stages (Graham & Perin, 2007):

1. Students are explicitly taught background knowledge needed to use a strategy successfully.
2. The strategy—as well as its purpose and benefits—is described and discussed.
3. The teacher models how to use the strategy.
4. Students memorize the steps of the strategy and any mnemonic associated with it.
5. The teacher supports or scaffolds student mastery of the strategy.
6. Students use the strategy with few or no supports.

Students are also taught a number of self-regulation skills including goal setting, self-monitoring, self-instruction, and self-reinforcement. These skills help students manage the writing strategies, the writing process, and their behavior during instruction.

Studies of SRSD in writing represent one of the most consistent efforts to explore the specific features of an instructional intervention, including systematic replications of research. Although previous reviews have been conducted on SRSD (Graham, 2006; Graham & Harris, 2003), the research base has not been evaluated using recently proposed quality indicators and standards for evidence-based practices in special education (Gersten et al., 2005; Horner et al., 2005). Examining the presence of quality indicators in research studies and considering only the high-quality studies that meet these criteria for methodological rigor allows reviewers to examine the impact of an intervention using only the most trustworthy sources of information. The current study analyzes the research evidence for SRSD in writing for students with LD or at risk for LD by evaluating (a) the presence of methodological quality indicators in research studies investigating this approach and (b) whether the bodies of group experimental and single-subject research meet standards for evidence-based practice proposed by Gersten et al. (2005) and Horner et al.

**METHOD**

Our research team carried out its work in four phases: (a) identifying intervention studies on SRSD in writing with students with LD or at risk for LD; (b) screening the studies to ensure they met inclusion criteria; (c) development, refinement, and application of a quality indicator rubric, based on published standards, for evaluating the methodological quality of the studies; and (d) application of published quality indicators and standards to determine whether the studies were of sufficient quality to deem SRSD an evidence-based practice. Each phase is described here in detail.

**Phase I: Identifying the Literature**

To conduct a thorough search of literature focused on SRSD in writing instruction, we completed a three-step process. First, we searched ArticleFirst, Educational Resources Information Center (ERIC), Google Scholar, OVID PsycINFO, and WorldCat electronic databases for studies conducted from January 1975 through December 2006. The start date was based on the year Congress passed the Education for All Handicapped Children Act of 1975 (Public Law 94-142). The following literature search terms were included: elementary education; learning disabilities; learning strategies; program effectiveness; secondary education; self-control; self-management; self-regulated, self-regulatory, special needs students; teaching methods; teaching models; writing difficulties; writing improvement; writing instruction; writing strategies; and writing skills. Electronic searches used multiple combinations of the literature search terms.
Second, we conducted an ancestral search using the reference lists from three secondary sources that focused specifically on research on writing instructional interventions for students with LD or students struggling with writing. These sources included: Baker et al. (2003), Gersten and Baker (2001), and Graham and Perin (2007). Third, we conducted a hand search of recent literature in the major journals of special, remedial, elementary, and secondary education. The following journals were searched through December 2006: American Journal of Education; Assessment for Effective Intervention; Cognition and Instruction; Educational Researcher; Educational Psychology; Elementary School Journal; Exceptional Children; Journal of Applied Behavior Analysis; Behavioral Disorders; Journal of Educational Psychology; Journal of Education and Behavior Statistics; Journal of Educational Research; Journal of Experimental Education; Journal of Experiential Education; Journal of Literacy Research; Journal of Negro Education; Journal of Special Education; Journal of Special Education Technology; Journal of Speech, Language, and Hearing Research; Learning Disability Quarterly; Memory and Cognition; Mental Retardation; Peabody Journal of Education; Reading Horizons; Reading Improvement; Reading Research and Instruction; Reading Research Quarterly; Reading Teacher; Remedial and Special Education; School Psychology Review; and Scientific Studies of Reading.

**PHASE II: SCREENING STUDIES FOR MINIMUM INCLUSION CRITERIA**

The search process previously described resulted in the identification of 49 articles pertaining to SRSD in writing. From these 49 studies, 21 (43%) met the following criteria for inclusion in our analysis: (a) publication in a peer-reviewed journal published in English between January 1975 and December 2006; (b) inclusion of students with LD or at risk for LD in kindergarten through Grade 12; (c) inclusion of SRSD as an independent variable in English language arts; (d) inclusion of a dependent measure of writing performance; and (e) use of an experimental, quasi-experimental, or single-subject design. We did not include dissertations. 

**PHASE III: DEVELOPMENT, REFINEMENT, AND APPLICATION OF A QUALITY INDICATOR RUBRIC**

After discussing the parameters of this special issue, our research team discussed an approach to evaluating the quality of each study. We determined that it would be very difficult to rate each study only for the presence or absence of each quality indicator outlined by Gersten et al. (2005) and Horner et al. (2005). Consequently, we created a rubric for both research designs (i.e., group design, single-subject) to evaluate the proposed quality indicators. Rubrics were designed jointly by the authorship team, reviewed, discussed, and revised. The final rubrics are presented in Chard, Ketterlin-Geller, Baker, Doabler, and Apichatabutra (2009; Figures 1 and 2). For each quality indicator, a 4-point rating system was used, with 1 being the lowest score and 4 being the highest score.

The development of the rubrics was an iterative process. We generated the initial rubrics to reflect the quality indicators and their components as described by Gersten et al. (2005) and Horner et al. (2005) as closely as possible. Specifically, for group experimental research, 10 components were categorized into four essential quality indicators defined by Gersten et al. (2005; see Table 1): (a) description of participants, (b) description and implementation of the intervention and comparison groups, (c) outcome measures used, and (d) data analytic techniques.

For single-subject research, components were organized into seven quality indicators defined by Horner et al. (2005; see Table 2): (a) description of participants and setting, (b) dependent variable, (c) independent variable, (d) baseline, (e) experimental control/internal validity, (f) external validity, and (g) social validity.

All authors then reviewed the rubrics to identify specific terms that were confusing or could lead to multiple interpretations. Once the rubrics had been initially revised, the research team practiced rating two articles that were not included in the final review. This practice rating allowed the team to discuss any ambiguous indicators or ratings and to further refine the rubrics for clarity and objectivity. After completion of the final rubrics, two independent reviewers rated the
### TABLE 1
Summary Scores of Self-Regulated Strategy Development Writing Interventions for Group Experimental and Quasi-Experimental Research

<table>
<thead>
<tr>
<th>Study</th>
<th>Description of Participants</th>
<th>Intervention/Comparison Conditions</th>
<th>Outcome Measures</th>
<th>Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>De La Paz &amp; Graham (1997a)</td>
<td>21 (3.5)</td>
<td>24 (4.0)</td>
<td>15 (3.75)</td>
<td>13 (3.25)</td>
</tr>
<tr>
<td>Graham, Harris, &amp; Mason (2005)</td>
<td>21 (3.5)</td>
<td>24 (4.0)</td>
<td>16 (4.0)</td>
<td>15 (3.75)</td>
</tr>
<tr>
<td>Harris, Graham, &amp; Mason (2006)</td>
<td>22 (3.67)</td>
<td>23 (3.83)</td>
<td>15 (3.75)</td>
<td>16 (4.0)</td>
</tr>
<tr>
<td>MacArthur, Schwartz, &amp; Graham (1991)</td>
<td>22 (3.5)</td>
<td>15 (2.5)</td>
<td>15 (3.75)</td>
<td>16 (4.0)</td>
</tr>
<tr>
<td>Sawyer, Graham, &amp; Harris (1992)</td>
<td>20 (3.33)</td>
<td>20 (3.33)</td>
<td>16 (4.0)</td>
<td>12 (3.0)</td>
</tr>
</tbody>
</table>

Note. Minimum summary score per essential quality indicator: \( a = 18, \quad b = 18, \quad c = 12, \quad d = 12 \). The mean score for the indicator is in parentheses. Mean scores below 3.0 do not meet the minimum quality indicator score.

Studies that met the initial screening criteria. The review team included two individuals each with more than 20 years of experience teaching and researching in the field of special education and an advanced doctoral student studying special education. Reviewers’ scores were aggregated across raters and across components (a number of specific components were rated for each broad quality indicator). A study met the overall quality indicator if it (a) received a minimum mean score across two reviewers of 3 or better averaged across the components for that specific quality indicator and (b) received no component score of 1 from either reviewer. A rating of 1, by one of the two reviewers, was automatically scored by a third reviewer. When the third reviewer determined that the component merited a rating higher than 1, that reviewer met with the two original reviewers. They discussed the score for the component in question and reached a consensus on the final rating. When the third reviewer agreed with the component score of 1, the score remained.

The cut-off score of 3, though arbitrary, was determined to be acceptable because quality indicators with a mean rating of 3 provided, on average, some evidence and/or description for the components of that quality indicator. The additional requirement that no component could receive a rating of 1 assured that at least some level of evidence and/or description was provided for each component of the quality indicators.

**Phase IV: Applying Published Quality Indicators and Standards**

According to Gersten et al. (2005), to be high quality, an experimental or quasi-experimental research study must (a) meet all but one of the Essential Quality Indicators and (b) demonstrate at least four of the Desirable Quality Indicators. An acceptable study must (a) meet all but one of the Essential Quality Indicators and (b) demonstrate at least one of the Desirable Quality Indicators. We applied the Essential Quality Indicators as outlined in the rubric in Chard et al. (2009; Figures 1 and 2) to the studies that employed experimental or quasi-experimental group designs. Once a study was determined to have met the Essential Quality Indicator criteria, it was reviewed by both reviewers to determine whether each of the Desirable Indicators was met. According to Horner et al. (2005), a study employing a single-subject design is high quality if it meets all of the methodological criteria outlined in the rubric in Chard et al.

Five research studies were evaluated that employed either an experimental or quasi-experimental research design. Sixteen single-subject research studies were analyzed for correspondence with the quality indicators identified by Horner et al. (2005). Interrater reliability of the independent ratings was calculated for both exact match agreement and in defining agreement as up to a 1-point discrepancy between the two reviewers. Reliability was calculated by dividing the number
| Study                                      | Participants/ Setting | Dependent Variable(s)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Danoff, Harris, &amp; Graham (1993)</td>
<td>22 (3.67)</td>
<td>34 (3.4)</td>
</tr>
<tr>
<td>De La Paz (1999)</td>
<td>24 (4.0)</td>
<td>38 (3.8)</td>
</tr>
<tr>
<td>De La Paz (2001)</td>
<td>22 (3.67)</td>
<td>35 (3.5)</td>
</tr>
<tr>
<td>De La Paz &amp; Graham (1997b)</td>
<td>16 (2.66)</td>
<td>38 (3.8)</td>
</tr>
<tr>
<td>Graham &amp; Harris (1989)</td>
<td>19 (3.17)</td>
<td>33 (3.3)</td>
</tr>
<tr>
<td>Graham &amp; MacArthur (1988)</td>
<td>20 (3.33)</td>
<td>35 (3.5)</td>
</tr>
<tr>
<td>Graham, MacArthur, Schwartz, &amp; Paige-Voth (1992)</td>
<td>19 (3.17)</td>
<td>35 (3.5)</td>
</tr>
<tr>
<td>Harris &amp; Graham (1985)</td>
<td>20 (3.33)</td>
<td>36 (3.6)</td>
</tr>
<tr>
<td>Lienemann, Graham, Leader-Janssen, &amp; Reid (2006)</td>
<td>22 (3.67)</td>
<td>34 (3.4)</td>
</tr>
<tr>
<td>Mason, Snyder, Sukhram, &amp; Kedem (2006)</td>
<td>17 (2.83)</td>
<td>36 (3.6)</td>
</tr>
<tr>
<td>Reid &amp; Lienemann (2006)</td>
<td>22 (3.67)</td>
<td>35 (3.5)</td>
</tr>
<tr>
<td>Saddler (2006)</td>
<td>20 (3.33)</td>
<td>35 (3.5)</td>
</tr>
<tr>
<td>Saddler, Moran, Graham, &amp; Harris (2004)</td>
<td>20 (3.33)</td>
<td>34 (3.4)</td>
</tr>
<tr>
<td>Sexton, Harris, &amp; Graham (1998)</td>
<td>20 (3.33)</td>
<td>38 (3.8)</td>
</tr>
<tr>
<td>Stoddard &amp; MacArthur (1993)</td>
<td>20 (3.33)</td>
<td>39 (3.9)</td>
</tr>
<tr>
<td>Troia, Graham, &amp; Harris (1999)</td>
<td>20 (3.33)</td>
<td>38 (3.8)</td>
</tr>
</tbody>
</table>

**Quality Indicators**

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>EC/Internal Validity</th>
<th>External Validity</th>
<th>Social Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danoff, Harris, &amp; Graham (1993)</td>
<td>14 (3.5)</td>
<td>23 (3.83)</td>
<td>8 (4.0)</td>
<td>31 (3.88)</td>
</tr>
<tr>
<td>De La Paz (1999)</td>
<td>24 (4.0)</td>
<td>15 (3.75)</td>
<td>23 (3.83)</td>
<td>31 (3.88)</td>
</tr>
<tr>
<td>De La Paz (2001)</td>
<td>19 (3.17)</td>
<td>14 (3.5)</td>
<td>23 (3.83)</td>
<td>31 (3.88)</td>
</tr>
<tr>
<td>De La Paz &amp; Graham (1997b)</td>
<td>23 (3.83)</td>
<td>7 (3.5)</td>
<td>29 (3.63)</td>
<td></td>
</tr>
<tr>
<td>Graham &amp; Harris (1989)</td>
<td>19 (3.17)</td>
<td>14 (3.5)</td>
<td>19 (3.17)</td>
<td>28 (3.50)</td>
</tr>
<tr>
<td>Graham &amp; MacArthur (1988)</td>
<td>18 (3.5)</td>
<td>14 (3.5)</td>
<td>21 (3.5)</td>
<td>27 (3.38)</td>
</tr>
<tr>
<td>Graham, MacArthur, Schwartz, &amp; Paige-Voth (1992)</td>
<td>17 (2.83)</td>
<td>13 (3.25)</td>
<td>19 (3.17)</td>
<td>28 (3.5)</td>
</tr>
<tr>
<td>Harris &amp; Graham (1985)</td>
<td>17 (2.83)</td>
<td>13 (3.25)</td>
<td>18 (3.0)</td>
<td>29 (3.63)</td>
</tr>
<tr>
<td>Lienemann, Graham, Leader-Janssen, &amp; Reid (2006)</td>
<td>24 (4.0)</td>
<td>8 (2.0)</td>
<td>20 (3.33)</td>
<td>30 (3.75)</td>
</tr>
<tr>
<td>Mason, Snyder, Sukhram, &amp; Kedem (2006)</td>
<td>22 (3.67)</td>
<td>10 (2.5)</td>
<td>20 (3.33)</td>
<td>27 (3.38)</td>
</tr>
<tr>
<td>Reid &amp; Lienemann (2006)</td>
<td>22 (3.33)</td>
<td>10 (2.5)</td>
<td>19 (3.17)</td>
<td>30 (3.75)</td>
</tr>
<tr>
<td>Saddler (2006)</td>
<td>20 (3.33)</td>
<td>12 (3.0)</td>
<td>21 (3.5)</td>
<td>29 (3.63)</td>
</tr>
<tr>
<td>Saddler, Moran, Graham, &amp; Harris (2004)</td>
<td>20 (3.33)</td>
<td>13 (3.25)</td>
<td>22 (3.67)</td>
<td>28 (3.5)</td>
</tr>
<tr>
<td>Sexton, Harris, &amp; Graham (1998)</td>
<td>20 (3.33)</td>
<td>13 (3.25)</td>
<td>24 (4.0)</td>
<td>29 (3.63)</td>
</tr>
<tr>
<td>Stoddard &amp; MacArthur (1993)</td>
<td>20 (3.33)</td>
<td>14 (3.5)</td>
<td>23 (3.83)</td>
<td>28 (3.5)</td>
</tr>
<tr>
<td>Troia, Graham, &amp; Harris (1999)</td>
<td>20 (3.33)</td>
<td>13 (3.25)</td>
<td>21 (3.5)</td>
<td>26 (3.25)</td>
</tr>
</tbody>
</table>

**Note.** Minimum summary score per quality indicator: *a* = 18, *b* = 30, *c* = 18, *d* = 12, *e* = 18, *f* = 6, *g* = 24. The mean score for the indicator is in parentheses. Mean scores below 3.0 do not meet the minimum quality indicator score.

EC = experimental control.
of agreements by the sum of agreements and disagreements. The reliability for exact agreements was 61.9% for single-subject studies and 76.7% for the group design studies. For matches that allowed a 1-point discrepancy, reliability was 96.4% for the single-subject studies and 93.3% for the group design studies. Interrater reliability on the presence of at least four of the desirable indicators was 100%.

For the group research studies, to be considered an evidence-based practice for students with or at risk for LD, there would need to be at least four acceptable studies, or two high-quality studies, that support SRSD in writing. Also the weighted effect size of these studies would need to be greater than 0 (Gersten et al., 2005). To determine if the weighted effect size was greater than 0, we first calculated a single effect size for each study based on the mean effect size across measures used to assess student writing performance (Gersten & Baker, 2001). The effect size calculated was Cohen's $d$, defined as the difference between the treatment and comparison group means divided by the pooled standard deviation (Cooper & Hedges, 1994). We calculated the posttest effect sizes by adjusting for pretest performance using the equation recommended by Wortman and Bryant (1985). To calculate the overall mean effect sizes across studies, we weighted the individual effects from each study by the sample size (Cooper & Hedges).

For single-subject studies, seven indicators (see Table 2) were used to evaluate the quality of each study. According to criteria proposed by Horner et al. (2005), a total of five single-subject studies would need to meet all of these quality indicators for the practice to be considered evidence based. In addition, these studies would need to have been conducted by at least three different researchers across at least three different geographical locations, and there would need to be at least 20 total participants across the studies.

RESULTS

QUALITY OF THE EXPERIMENTAL AND QUASI-EXPERIMENTAL RESEARCH STUDIES

The five group research studies were evaluated for the description of participants, the description and implementation of the intervention and comparison conditions, the outcome measures used, and data analytic techniques. For its description of the participants, the studies were evaluated for the quality of student demographic information including disability diagnoses procedures and comparability of students across conditions. Information about the interventionists or teachers implementing the intervention was evaluated along with the comparability of their characteristics and/or credentials across conditions. All five studies met the minimum criteria of an average score of 3 across all subindicators with no 1-point scores.

Next, studies were evaluated for their description and implementation of the intervention, fidelity of implementation procedures, and description of the comparison condition. Four of the five studies met or exceeded the minimum requirements for rigorous research in this methodological category (i.e., a mean score greater than 3). The study that did not meet the minimum criteria (MacArthur et al., 1991) failed to describe the process for determining fidelity of implementation and, therefore, received a rating of 1 which meant it could not meet the criteria for this category. It also fell below a mean score of 3 for this indicator.

All five studies met the minimum criteria for rigorous research methodology for documentation of outcome measures. All studies documented appropriate data collection times and used multiple dependent measures or measures of generalized performance to document the effects of the independent variable. The five studies also employed appropriate data analytic techniques aligned with the research question, and the data were analyzed at the appropriate unit of statistical analysis.

In summary, based on our evaluation of essential quality indicators of group experimental and quasi-experimental intervention research, all five of the SRSD group studies met the standard proposed by Gersten et al. (2005) for a high-quality study in relation to the essential quality indicators (i.e., all but one essential quality indicator adequately addressed). Gersten et al. suggested that a high-quality study should also meet at least four of the desirable quality indicators, and an acceptable study should meet at least one
of the desirable quality indicators. We documented the presence of Desirable Indicators of quality research to determine whether each SRSD group intervention study met criteria to be considered acceptable or high quality. In our analysis, all five studies provided adequate evidence of at least four desirable indicators by (a) documenting insignificant attrition of participants across conditions, (b) giving considerable attention to describing the nature of the comparison condition, (c) making efforts to measure outcomes beyond the immediate posttest, and (d) reporting the research coherently. As such, the research reported in all five studies should be considered high quality.

The effect sizes for each of the five SRSD group studies ranged from +0.80 to +1.85. The average weighted effect size was +1.22 and the 95% confidence interval around this effect size ranged from a low of +0.92 to a high of +1.53. Because this confidence interval does not include 0, the set of high-quality group design SRSD studies met the Gersten et al. (2005) criteria for being evidence based for students with and at risk for LD.

QUALITY OF THE SINGLE-SUBJECT RESEARCH STUDIES

The 16 single-subject studies were reviewed for their discussion of the description of participants and setting: the sample characteristics, process for selecting participants, and description of the critical features of the physical setting. All but two studies met the minimum requirements of acceptability for this category. The studies that did not meet this quality indicator (i.e., De La Paz & Graham, 1997b; Mason, Snyder, Sukhram, & Kedem, 2006) did not provide sufficient details about the physical setting, including where the instruction took place and the context of instruction.

Each study was evaluated for its treatment of the dependent variable including documentation of the dependent variable, measurement procedures including technical adequacy, and frequency and reliability of data collection. All 16 studies met the overall minimum criteria for this category by receiving a mean score of at least 3 across these subindicators.

Documentation of the overall description, manipulation, and fidelity of implementation of the independent variable was evaluated to determine if each research study provided evidence consistent with the criteria for rigorous research. Thirteen of the 16 research studies met the minimum criteria for acceptability in this category. Two studies (Graham, MacArthur, Schwartz, & Page-Voth, 1992; Harris & Graham, 1985) failed to achieve the mean score of 3 and also received a score of 1 on the quality indicator related to fidelity of implementation. A third study (Graham & MacArthur, 1988) earned the necessary minimum mean score but received a score of 1 on the quality indicator related to fidelity of implementation. The reviewers determined that these three studies did not provide sufficient evidence that fidelity was measured or that the manner in which it was measured did not ensure that the results of the studies could be attributed to the implementation of the independent variable.

The studies were evaluated to determine whether they met minimum criteria for documentation of the description of the baseline conditions. Thirteen of the 16 studies met the minimum criteria. Three studies (Lienemann, Graham, Leader-Janssen, & Reid, 2006; Mason et al., 2006; Reid & Lienemann, 2006) did not meet the minimum criteria. In the Lienemann et al. study, student performance was not sufficiently consistent to predict future performance or to detect a change in level or trend upon introduction of the independent variable. Also, in all three studies, the description of the baseline condition was not sufficient to determine its distinctiveness from the SRSD condition. Consequently, it was difficult to determine unequivocally that the changes that occurred during the experimental condition were related to the independent variable.

Each study was evaluated to determine whether it met minimum criteria for demonstrating experimental control by systematically manipulating the independent variable while documenting concomitant changes in the dependent variable(s). In all cases, the studies employing single-subject designs demonstrated experimental control both by staggering the introduction of the independent variable and documenting changes in trend and level of the dependent variable or by
manipulating the independent variable across different observation periods.

To determine the adequacy of the external validity of the research documented in each study, we reviewed each study's description of the potential generalizability of results. Each study was evaluated to determine whether there were efforts to replicate the findings across participants, settings, or materials that were in all 16 studies.

Fifteen of the 16 research studies provided adequate documentation to justify the social validity of the research. The studies were reviewed across four dimensions: (a) importance of the dependent variable, (b) importance of the magnitude of change in the dependent variable, (c) practicality and cost effectiveness of the independent variable, and (d) nature of implementation of the independent variable. The study that did not meet minimum criteria (Troia, Graham, & Harris, 1999) was rated 1 on the nature of the implementation of the independent variable because the intervention was not implemented by a certified teacher or this consideration was not clearly articulated in the description of implementation.

Across the 16 studies, 9 achieved a mean score of 3 or above in each of the seven quality criteria categories, and none of the components were rated as 1. In addition, these studies were conducted by at least three researchers in at least three different geographical locations, and the number of student participants across the studies exceeded 20. According to Horner et al. (2005) at least five studies are needed that meet these criteria in order for a practice to be considered evidence based. Thus, based on the single-subject research studies we reviewed, SRSD would be considered to be an evidence-based practice.

**DISCUSSION**

This study evaluated the quality of published research on SRSD, a comprehensive approach to writing instruction, to determine if (a) the research meets standards of high quality and rigor as proposed by Horner et al. (2005) and Gersten et al. (2005) and (b) SRSD could be considered an evidence-based practice, based on the standards proposed by Horner et al. and Gersten et al. (2005).
of implementation. Consequently, these studies did not meet the criteria for the indicator targeting the independent variable. Both met the minimum criteria for the remaining six quality indicators. De La Paz and Graham (1997b) and Mason et al. (2006) did not meet the minimum criteria for describing the participants and settings. In both cases, the reviewers rated the studies particularly low on their description of the setting in which the intervention was delivered. Three studies (Lienemann et al., 2006; Mason et al.; Reid & Lienemann, 2006) also met the minimum score for six of the seven indicators, but received less than the minimum score on the indicator related to the quality of the description of the baseline condition. The final study (Troia, Graham, & Harris, 1999) received a score of 1 on the criteria related to the nature of the implementation because the intervention was delivered by the first author of the study rather than by typical agents in the school. Thus, applying the criteria proposed by Horner et al., these seven studies do not meet rigorous methodological standards.

**SRSD as an Evidence-Based Practice**

**Group Design Research Studies.** All five group studies adequately addressed each of the essential quality indicators proposed by Gersten et al. (2005) for high-quality research. All five also met at least four desirable quality indicators: (a) insignificant attrition of participants across conditions, (b) considerable attention given to describing the nature of comparison condition, (c) efforts made to measure outcomes beyond the immediate posttest, and (d) coherent reporting of the research. Based on the criteria for essential and desirable quality indicators, each of the SRSD group studies met standards for high-quality research.

For a practice to be considered evidence based, Gersten et al. (2005) recommended that at least two high-quality group studies support the practice and that the effect size, weighted by the sample size of the individual studies that are included in the effect size calculation, is significantly greater than 0 (i.e., the 95% confidence interval of the weighted mean effect size does not include 0). The weighted mean effect size in the high-quality SRSD group research studies was +1.22, with a 95% confidence interval that did not include 0. Therefore, the set of SRSD group research studies met standards for an evidence-based practice for students with and at risk for LD.

**Single-Subject Research Studies.** For a practice investigated through single-subject research designs to be considered evidence based, Horner et al. (2005) suggests the following standards be met: (a) The practice must be supported by a minimum of five studies, published in peer-reviewed journals that meet minimally acceptable methodological criteria and document experimental control; (b) the studies must be conducted by at least three different researchers across at least three different geographic locations; and (c) the five or more studies include a total of at least 20 participants.

Nine studies reviewed in the investigation qualify as high-quality single-subject research. These studies have been conducted by at least three different research teams and in three different geographic locations. The total sample size was greater than 20 participants in these studies. Also, criteria for high-quality single-subject studies included experimenter control in the implementation of the independent variable, which resulted in determinations of the positive impact of the SRSD when implemented. Consequently, SRSD meets the standards proposed by Horner et al. (2005) for an evidence-based practice on the basis of single-subject research for students with and at risk for LD.

**Study Selection**

One of our inclusion criteria was that the study had to be published in a peer-reviewed journal. Three considerations are important in this regard. First, we reasoned that selecting studies in peer-reviewed journals should have given us the best opportunity for evaluating the highest quality studies of SRSD. Our interest was in rating the best quality research, and accessing studies published in peer-reviewed journals provided the best opportunity to identify high-quality studies, given that they had been vetted through blind peer review. Second, including only studies published in journals enabled us to examine the standards that journals in the fields of general and special education
apply for methodologically acceptable intervention studies. Finally, studies published in professional journals contain evidence readily available to other researchers and practitioners in our field. In evaluating these studies, we analyzed evidence that others in the field could be expected to investigate when analyzing the impact of SRSD.

Implications for Research and Practice

Although the frameworks posed by Horner et al. (2005) and Gersten et al. (2005) are specific in terms of what should be included in high-quality single-subject and group design studies, they do not specify how to measure each quality indicator. It would have been possible to develop some type of dichotomous rating scheme for each indicator, which we considered. However, we opted for a different approach. For both the group design and single-subject studies, we decided on a 4-point rating scale after much discussion and practice with different types of formats. Ultimately, we felt that the presence of essential quality indicators was best measured on a continuum, rather than dichotomously. The quality indicators were often partially present, rather than completely absent or fully present, in the studies we were reviewing as well as in the intervention research studies with which we are familiar in special and general education.

The reliability of our ratings should be considered in the context of our decision to develop a multiple-point scale. Our ability to agree exactly was 76.7% for group studies and 61.9% for the single-subject studies. In terms of traditional standards of reliability, this may seem low. However, in the context of quality ratings of this sort, such as classroom observation research (Baker, Gersten, Haager, & Dingle, 2006; Stanovich & Jordan, 1998), this degree of correspondence may be more acceptable. When agreements were defined as either exact agreements or agreements within 1 point of each other, our reliabilities improved to 93.3% for the group research studies and 96.4% for the single-subject research studies. We do realize that on a 4-point scale, a 1-point discrepancy may still be too high, but we think that in this context it represents an important degree of correspondence that is lost when only exact match agreement data are considered as an acceptable standard.

We have three comments about how the reliability ratings of the quality of intervention studies might be improved. The first is that better measurement approaches than we used may be developed and implemented. The value of the criteria posed by Horner et al. (2005) and Gersten et al. (2005) is that there are clear components of high-quality intervention research that should be considered in instrument development. Other researchers can construct measurement instruments, or build on the instruments we have developed, and use them to evaluate intervention research. An important goal will be to demonstrate the reliability and validity of these instruments. A second way to improve the ratings of research quality may be to make the standards posed by Horner et al. and Gersten et al. (2005) more commonly known and accepted by research teams submitting their work for review. Common research standards for intervention research are becoming more acknowledged and widespread, and we would like to see this trend continue. If the indicators proposed by Horner et al. and Gersten et al. (2005) did become more widely adopted and used by journal editors in reviewing studies for publication, for example, then researchers would have a clear incentive to conduct their research in a way that was aligned with very specific aspects of research quality. As research interventions become more complex, more widespread use of common quality standards like the standards posed by Gersten et al. (2005) and Horner et al. become even more important.

Our third comment is that it may be increasingly beneficial to use Web-based links in manuscripts that include important information about methods and data analysis procedures that frequently cannot be included in the print copies of journals because of space limitations. These links could include additional information for analysis of research quality. In intervention research, for example, it is frequently a challenge to describe all aspects of the intervention procedures that might be necessary for replication, particularly if the intervention is complex. Because the print copy of journal space is expensive, detailed descriptions of interventions and treatment procedures are commonly abbreviated to cover just the
essentials. It would be possible, however, to include Web-based links within the journal to provide fuller explanations of procedures for implementation or data analytic approaches, which might be necessary to provide a full rendering of the quality of an intervention study. The important point is that as intervention studies are reviewed for quality, authors and reviewers should have reasonable opportunities to provide and access full information about the study, particularly if criteria addressing key aspects of quality are not contained in the printed copy of the study.

The research base of SRSD in writing also has implications for practice. SRSD provides a comprehensive approach to writing instruction with a clear set of procedures and steps to follow to implement the practice the way it has been used in research studies. Students are explicitly taught strategies for planning, drafting, and revising text (Graham & Perin, 2007). They are also taught a number of self-regulation skills including goal setting, self-monitoring, self-instruction, and self-reinforcement. The evidence suggests that if the approach is implemented with fidelity, schools can expect to see a significant improvement in students' writing. Based on the studies we reviewed, the high-quality research has been conducted primarily with students with LD or at risk for LD. The moderate to strong improvements in the writing achievement schools might expect should be anticipated primarily with these types of students. Also, most of the studies have been conducted in late elementary and middle school grades, and procedurally it will be most straightforward for districts and schools to consider the implementation of SRSD in writing in these grades.

It may be useful, however, to consider reasonable uses of SRSD in ways that extend how the practice has been used specifically in the research studies we reviewed. The dismal writing achievement of students in this country across different student groups and grade levels suggests that writing interventions implemented across the range of students in typical general education settings could be a valuable investment. In the context of SRSD specifically, there is nothing in the nature of the approach that would suggest it might inhibit the writing achievement of any particular group of students (e.g., students who are not at risk for LD or expected to experience writing problems). The focus of SRSD is on providing teachers with a comprehensive approach to writing instruction that can help students write more effectively and coherently across writing genres. On its surface, it seems very reasonable to expect that the majority of students in elementary and middle school settings could benefit from this approach.

Another way to consider the use of SRSD is in a multitiered approach, in which a core program, in this case in writing, is used as the basis of Tier 1 instruction. Depending on how students receiving Tier 1 instruction respond, the intensity of instruction for some students may need to be increased systematically within the general education setting before students are referred, evaluated, and placed in special education. Given the paucity of evidence-based writing programs, a strong argument could be made that SRSD could serve as a core or Tier 1 writing program. For students who struggle to develop effective writing skills—despite strong instruction using SRSD in Tier 1—further increases in the intensity of writing instruction could be implemented within Tier 1 and in subsequent tiers. Because SRSD does have strong evidence of effectiveness for students at risk for LD or with LD, the approach could serve as the basis of writing instruction with increased intensity of writing instruction for students who continue to struggle.

The SRSD Research Team
SRSD research has been conducted by researchers who share a common tradition with this particular intervention, including active collaboration with Graham and Harris, who conceptualized and developed SRSD. The cohesive nature of the research group has two strengths, one of which we believe is reflected directly in our analysis. Our high ratings of this research on quality indicators in both group design and single-subject design studies reflect a shared understanding among SRSD researchers to conduct high-quality research.

A second advantage, not readily apparent from our analysis, is that the core set of researchers has constructed a cohesive line of research. This is reflected in the developmental
nature of SRSD studies, in which there is a consistent progression from investigations of the overall impact of SRSD to investigations of more detailed or subtle aspects of SRSD that build explicitly on the findings from previous studies. For example, Graham et al. (2005) and Harris et al. (2006) investigated the impact of SRSD with two intervention groups and a control group. The two intervention groups differed in that one group also included a component with explicit peer support. This builds on the tradition of SRSD research by incorporating a theoretically driven and empirically supported component to support use, generalization, and maintenance of the writing strategies. This common commitment to iteratively refining the body of empirical studies examining SRSD illustrates the power of research on an instructional intervention that is well conceived, grounded in a strong theoretical foundation, and carefully implemented (Harris, 1990).

A limitation to a cohesive group of researchers conducting the research on SRSD is that the impact of SRSD when implemented by other researchers not associated with the core research team has not been investigated. Particularly with a complex intervention such as SRSD, it is important to know what the replication effects are when researchers not connected with the original development team study the impact of the intervention. The importance of studying the effectiveness across multiple research teams is reflected in Horner et al.’s (2005) standards for determining if a practice is evidence based (i.e., studies must be conducted by at least three different researchers). Although SRSD in writing has been conducted by a number of researchers, it would be valuable if additional researchers outside of the SRSD tradition conducted serious research on SRSD. In addition to providing important information about the effectiveness of the intervention when implemented under conditions of reasonably high experimental control, additional replications of SRSD would provide important information about the potential effectiveness of the intervention if it were scaled up in a major way.

For example, it would be very interesting to examine the impact of SRSD in a high-quality research study when a school district or a consortium of schools made the effort to evaluate the impact of SRSD when it is implemented using their own professional development resources, district trainers, and so forth. The latest NAEP data (2007, http://nationsreportcard.gov/writing_2007/w0001.asp) with 8th and 10th graders (including students with disabilities), and the 2002 NAEP data with 4th, 8th, and 10th graders suggests that comprehensive approaches to writing instruction should be more broadly adopted by schools. SRSD offers credible evidence that it could significantly improve the writing achievement of students, including students with disabilities, if implemented with fidelity. We see this opportunity to expand the research teams that implement SRSD and variations of it as an exciting opportunity, given the evidence accumulated thus far regarding the quality of the SRSD intervention.

REFERENCES


De La Paz, S., & Graham, S. (1997a). Effects of dictation and advanced planning instruction on the com-
posing of students with writing and learning problems. *Journal of Educational Psychology, 89*(2), 203–222.


ABOUT THE AUTHORS

**SCOTT K. BAKER** (CEC OR Federation), Director, Pacific Institutes for Research, and Associate Director of the Center on Teaching and Learning at the University of Oregon, Eugene. **DAVID J. CHARD** (CEC TX Federation), Leon Simmons Endowed Dean, Annette Caldwell Simmons School of Education and Human Development, Southern Methodist University, Dallas, Texas. **LEANNE R. KETTERLIN-GELLER** (CEC OR Federation), Assistant Professor, Educational Leadership; **CHANISA APICHATABUTRA**, Doctoral Candidate; and **CHRISTIAN DOABLER** (CEC OR Federation), Doctoral Candidate, Department of Special Education and Clinical Sciences, College of Education, University of Oregon, Eugene.

Address correspondence to Scott Baker, Pacific Institutes for Research, 1600 Millrace Drive, Suite 109, Eugene OR 97403-1995 (e-mail: sbaker@uoregon.edu).

Manuscript received September 2007; accepted May 2008.