Interpreting the Efficacy of Social Skills Training:

Has the Picture Changed?

Kenneth A. Kavale and Mark P. Mostert

Regent University

Under review: Behavioral Disorders

Kenneth A. Kavale, PhD
Professor, Special Education
Regent University
757.226.4750
Fax 757.226.4247
Email: kkavale@regent.edu
1000 Regent University Drive
Virginia Beach, VA 23464-9800
Abstract

The efficacy of social skills training (SST) has been difficult to establish. Historically, SST has been viewed as a beneficial intervention for some children and youth with EBD in some circumstances. There is, however, an evolving position suggesting increasing efficacy for SST. A notable source for this new position are three reviews of meta-analyses investigating the efficacy of SST which reveal a shift from an earlier, cautious position about SST efficacy to one that now claims SST to be a more widely effective intervention. We argue that the change in position about SST efficacy is not warranted and is primarily the result of a variety of interpretive problems about how SST efficacy might best be gauged. We conclude that SST remains an intervention demonstrating only limited efficacy.
Interpreting the Efficacy of Social Skills Training:

Has the Picture Changed?

Since the 1980s, the efficacy of social skills training (SST) has often been contentiously debated. Although widely used because of its intuitive appeal, empirical support has been equivocal (Kavale & Mostert, 2004; McIntosh, Vaughn, & Zaragoza, 1991; Sridhar & Vaughn, 2001). The equivocal support makes it possible to question whether or not SST should be included in intervention programs. Although a popular treatment option, the question remains: Is it possible to change behavior to the point where social competence is significantly enhanced? Additionally, Maag (2005) pointed out a number of problems with SST research that require resolution before SST can be endorsed.

With the rapid expansion of the SST research base, research synthesis techniques, usually meta-analyses, have been used to gain greater insight into the efficacy of SST. The outcomes reveal that SST generally produces small to medium effects according to Cohen’s (1998) criteria. In turn, SST meta-analyses have been combined (mega-analyses) in an attempt to reach more definitive conclusions. For example, a recent mega-analyses by Gresham, Cook, Crews, and Kern (2004) concluded that, “Overall, SST is an effective and beneficial intervention for students with or at risk for EBD based on the reviewed meta-analyses” (p. 43). When compared to earlier mega-analytic conclusions where SST was “a relatively weak intervention strategy” (Gresham, 1998, p. 19) with “weak effects” (Gresham, Sugai, & Horner, 2001, p. 331), the recent positive evaluation stands in sharp contrast. Have research findings revealed greater positive effects for SST since 1998? Do
recent research findings provide sufficient justification for a more favorable view of SST?

Because the three mega-analyses offered by Gresham and colleagues (i.e., Gresham, 1998; Gresham et al., 2001; Gresham et al., 2004) represent the most long-standing and most often cited efforts evaluating the efficacy of SST, it is important to determine whether or not the evidence warrants the change to a more positive view of SST. The purpose of this paper is to indicate substantive difficulties in the three reviews that attenuate decisions about the efficacy of SST and, ultimately, demonstrate that SST continues to produce only modest effects.

Overview of Gresham and Colleagues’ Changing Picture of SST Efficacy

Meta-analysis brings a rigorous and systematic framework to the research synthesis process (Kavale, 2001). Similarly, when meta-analyses themselves are combined, a rigorous and systematic approach is also necessary, particularly in the crucial areas of study selection (What meta-analyses to include or exclude in the review) and interpreting results (Can the findings be properly considered in evaluating SST efficacy?).


Study selection. Gresham (1998) considered the possibility of “rebuilding” SST training because its effects were deemed weak, modest, unimpressive, and small. The conclusion about limited positive effects for SST was based on the findings from three SST meta-analyses (i.e., Beelmann, Pfingsten, & Losel, 1994; Kavale, Mathur, Forness, Rutherford, & Quinn, 1997; Schneider, 1992) “conducted in the 1990s” (p. 20). With no description about how the three meta-analyses were located or criteria for inclusion in the
mega-analysis, it is impossible to determine whether another meta-analysis conducted during the 1990s (i.e., Forness & Kavale, 1996) was either not located or not included according to selection criteria. The omission is notable since the Forness and Kavale meta-analysis was used in a later SST mega-analysis (see Gresham et al., 2001). Similarly, another meta-analysis might have been included were it not for the “1990s” criteria (i.e., Schneider & Byrne, 1985) which also appeared in a later meta-analysis (see Gresham et al., 2004).

Because the selection of meta-analyses for Gresham (1998) appears somewhat arbitrary, it seems that the statement “depending on which meta-analysis or narrative review of the SST literature one reads, different conclusions can be drawn regarding its efficacy” (p. 19) is indeed valid. With potentially significant information omitted, it is difficult to have confidence in Gresham’s (1998) conclusions about SST efficacy (see Table 1).

**Interpreting findings.** Gresham’s (1998) mega-analysis included findings from three meta-analyses (Beelman et al., 1994; Kavale et al., 1997; Schneider, 1992) that “produced effect sizes ranging from .20 to .47” (p. 20). The reported ES range, however, does not appear to include all three meta-analyses. The Schneider (1992) meta-analysis, with a mean ES of .89, seems to have been omitted; the proper ES range should be .20 to .89.

The Schneider (1992) meta-analysis also seems to have been omitted from the mean ES calculation. The reported mean ES of .35 can only be explained by combining Kavale et al. (1997, ES = .20) and Beelman et al. (1994, ES = .47). Gresham (1998) provided no explanation for excluding the Schneider meta-analysis whose ES of .89 (a
“large” ES according to Cohen’s [1988] guidelines for interpretation) would have
produced a weighted mean ES of .62, and the possibility of an interpretation different
from that of SST being a “relatively weak intervention strategy” (p. 19).


Study selection. The Gresham, et al. (2001) mega-analysis included “a total of
six meta-analyses [that] have been conducted using students with or at risk for high
incidence disabilities” (p. 336). Gresham et al. defined “high incidence disabilities”
(HID) as those students with learning disability (LD), mental retardation (MR), emotional
disturbance (ED), and attention deficit/hyperactivity disorder (ADHD). Upon
examination, however, one meta-analysis (i.e., Denham & Almeida, 1987) does not
include a sample that can be appropriately classified under the HID designation. The
reason for its inclusion is not explained.

Gresham et al. (2001) included two meta-analyses from the 1998 effort (i.e.,
Beelmann et al., 1994; Schneider, 1992), omitted Kavale et al. (1997), and added four
new meta-analyses (Coleman, Wheeler, & Webber, 1993; Denham & Almeida, 1987;
Forness & Kavale, 1996 [erroneously reported as 1999]; Mathur, Kavale, Quinn, Forness,
& Rutherford, 1998). This compilation of meta-analyses was not accompanied by
selection criteria making it difficult to explain the choices made. For example, two meta-
analyses (i.e., Schneider & Byrne, 1985; Quinn, Kavale, Mathur, Rutherford, & Forness,
1999), although seemingly appropriate, were not included. Additionally, although two
meta-analyses (i.e., Beelman et al., 1994; Schneider, 1992) were previously described as
“relatively modest” and “unimpressive,” three years later Gresham et al. suggested that
they now possess “substantially stronger effect sizes” (p. 336). Finally, the Mathur et al.
(1998) meta-analysis included only single-subject research making the statement that “Mathur et al. analyzed 35 group and 64 single-case design studies” (p. 336) incorrect. The Mathur et al. study would not provide a parametric ES appropriate for a mega-analysis.

Interpreting findings. Although six meta-analyses are listed, Gresham et al. (2001) reported “an average effect size across four meta-analyses of .437” [italics added] (p. 337). No explanation is provided for omitting two meta-analyses in calculating the mean ES. In one case, the reason is clear: Coleman et al. (1993) is not a meta-analysis but rather a narrative review containing no ES data. The reason for the exclusion of Denham and Almeida (1987) is less clear because it is a meta-analysis. Although not including an HID sample, Gresham et al. did not exclude this study on that basis and, consequently, Denham and Almeida should be part of the ES calculation.

The actual calculation of the mean ES from the four meta-analyses raises questions about the validity of the .437 figure. The “ES” (.20) presumably reported from Mathur et al. (1998) is actually the percent of nonoverlapping data points (PND) used to calculate effects in single case designs (PND = 62%). The PND is thus a different metric inform and function and, because it is not a parametric ES, is inappropriate in calculating a mega ES across meta-analyses. Perhaps the source of “.20” is actually the ES reported earlier by Kavale et al. (1997). If this is the case, then Kavale et al. probably should not have been excluded from the Gresham et al. review.

The four meta-analyses used by Gresham et al. (2001) appear to be the three used in Gresham (1998) plus one conducted by Forness and Kavale (1996). If these were the meta-analyses actually included, the weighted mean ES for Gresham et al. would be .52.
How was the Gresham et al. mean value of .437 calculated? Was the ES of .437 somehow calculated from the “six” meta-analyses listed? Was the same calculation error from 1998 carried through to 2001? Assuming our analysis of the meta-analyses actually used is correct, the salient question is how the addition of a mean ES of .21 (i.e., Forness & Kavale, 1996), a value smaller than the 1998 reported mean ES of .35, was able to increase the mean ES of Gresham et al. to .437.

There are additional interpretive problems. For example, the mean ES of .437 was interpreted to mean “that SST can improve students’ social competence” (p. 337) while the earlier Gresham (1998) review with only a slightly lower mean ES of .35 was deemed “unimpressive” (p. 20). The difference between the two mean ESs (.437 and .35) would amount to a modest 3% (14% to 17%) increase in positive social skill functioning, a finding which does not appear to warrant an endorsement of SST. Actually, compared to what the Gresham (1998) mean ES should have been (.62), the Gresham et al. (2001) findings (.437) actually demonstrate a 6% loss in SST efficacy. Even with what we believe to be the correct ES (.52), there would still be an overall 3% loss in SST efficacy. The proper conclusion, we think, is that SST may improve social competence but not nearly as much as suggested by Gresham et al.


*Study selection.* The most recent mega-analysis (Gresham et al., 2004) included six studies representing “a comprehensive examination of all the group-based meta-analyses performed on social skills training” (p. 33). The selection criteria included the stipulation that meta-analyses had to be published between 1987 and 2003, but one meta-analysis (i.e., Schneider & Byrne, 1985) was included despite the stipulated time
frame. The lack of further selection criteria makes it difficult to judge the validity of either adding or omitting particular meta-analyses. Although some meta-analyses are properly omitted (i.e., Coleman et al., 1993; Mathur et al., 1998), it remains difficult to determine whether the included meta-analyses represent a summative literature base that can be used to achieve the cumulative knowledge necessary for judging the efficacy of SST.

In describing the scope of the review, Gresham et al. (2004) suggested that “These meta-analyses focused on children and youth with behavioral difficulties, involved approximately 338 studies, and included more than 25,000 children between 3-18 years of age” (p. 36), but several difficulties should be noted. First, although Gresham et al. (2004) listed six meta-analyses, only five appear to have actually been used in their review. The meta-analysis by Quinn, Kavale, Mathur, Rutherford, and Forness (1999) appears to have been ignored. An examination of its title seems to suggest that it meets the purpose stipulated by Gresham et al. Second, although the subject n across meta-analyses appears impressive, one meta-analysis contributed a disproportionate share (the 16,623 subjects reported by Losel & Beelman, 2003) while four of the meta-analyses (i.e., Ang & Hughes, 2002; Beelmann et al., 1994; Schneider, 1992; Schneider & Byrne, 1995) actually did not report a subject n.

Interpreting findings. At a later point, Gresham et al. explained why they used five, not six, meta-analyses. The Quinn et al. (1999) meta-analysis was eliminated because it possessed a “curiously low effect size” (p. 43). Such an idiosyncratic exclusion criterion is not justified and may interfere with an authentic evaluation of SST. The findings of Quinn et al. (ES = .199) may be inconvenient for interpreting SST as an
“impressive” and “beneficial” intervention, but needs to be included to assure an unencumbered and unbiased evaluation of the efficacy of SST.

Gresham et al. (2004) reveals additional interpretive difficulties. The reported mean ES of .61 that was initially, “Based on the above five meta-analyses” (p. 38, i.e., all but Quinn et al., 1999), but later it is indicated that, “the six meta-analyses reviewed showed” (p. 43) a mean ES of .61. We question how the Quinn et al. meta-analysis with its “curiously” low ES (the reason for its exclusion) was able to produce the same mean ES (.61) when it was included after being excluded. Our calculation with six meta-analyses shows a mean ES of .56, thereby making it difficult to discern how “These results are even more impressive if one eliminates the curiously low effect size . . . reported by Quinn et al. (1999)” (p. 43).

In reality, the Gresham et al. (2004) findings do not substantially differ from the ES (.62) that should have been found in Gresham (1998). The ES (.56) we calculated from Gresham et al. (2004) would produce a 21 percentile rank gain, which is comparable to the 23 percentile rank gain that would be associated with the corrected ES (.62) from the 1998 mega-analysis. The absence of an upward trend in SST efficacy is apparent when “new” and “old” meta-analyses are compared. The three “new” meta-analyses (i.e., Ang & Hughes, 2002; Losel & Beelmann, 2003; Schneider & Byrne, 1985) produce a mean ES of .51, while the three “old” (i.e., previously used) meta-analyses (i.e., Beelmann et al., 1994; Quinn et al., 1999; Schneider, 1992) produce a mean ES of .62. Thus, the latest meta-analyses included in Gresham et al. actually reveal some regression in SST efficacy and maintain effects of SST at no more than a “medium” level (see Cohen, 1988).
In sum, the meta-analytic evidence appears not to have changed significantly since the original Gresham (1998) review. We see no apparent trend indicating increased efficacy. It is difficult, therefore, to endorse the Gresham et al. (2004) conclusion that, “Overall, SST is an effective and essential part of a comprehensive intervention program for students with EBD” (p. 32).

The Binomial Effect Size Display

An ES may be interpreted in a number of ways, but most of these explanations offer an abstract statistical rather than a practical rendering (see Kavale, 2001). We, therefore, endorse the decision of Gresham et al. (2004) to use the Binomial Effect Size Display (BESD) (Rosenthal & Rubin, 1982) for interpreting their findings. The BESD, based on converting ES to $r$, provides an answer to the important question: What is the effect on the success rate of using treatment X? For the Gresham et al. mega-analysis, the reported ES of .61 can be represented as a contrast between a SST treatment group with a success rate of 65% versus a comparison (control) group with a success rate of 35%. The BESD interpretation showed that almost two out of three children receiving SST demonstrate improvement. Gresham et al., however, appear to have misinterpreted the BESD: the 65% stipulated by Gresham et al. is not the “degree of improvement” (p. 43) but rather the percentage of children who will demonstrate some improvement (degree unknown).

As an index of practical significance, the BESD is dependent on the context of interpretation. A correlation coefficient (e.g., $r = .20$) is often underestimated because the usual interpretation indicates that it accounts for “only” 4% of the variance. The same correlation coefficient (i.e., $r = .20$) with a BESD interpretation indicates an increase in
success rate from 40% to 60%. If this BESD were associated with a medical study, for example, then a decrease in mortality rate from 60% to 40% is practically very significant. In fact, if an outcome is unencumbered (e.g., dead vs. alive) an r of .10 assumes practical importance. Conceptually, however, SST is not “life and death,” and assessing its practical significance is far less obvious. For example, across 156 meta-analyses of psychological, educational, and behavioral treatments, Lipsey and Wilson (1993) found a grand mean (mega) ES of .47 which translates into a 62% success rate. With the recalculated mean ES of .56 (i.e., with all six meta-analyses included), the Gresham et al. (2004) findings translate into a 64% success rate suggesting that SST is probably no better or worse than many other treatments. Thus, SST appears to produce some benefits in some instances; a perspective that holds true for many special education interventions.

Exclusion of the Quinn et al. (1999) Meta-analysis

Gresham et al. (2004) provided several reasons for not including the Quinn et al. (1999) meta-analysis. Its exclusion, however, eliminates potentially significant information. The reported ES (.199) should not be prejudged as “curiously low” but rather as a reflection of “what the research says” about SST for students with EBD. The Forness and Kavale (1996) meta-analysis investigating SST for students with LD found a similar ES (.21) but was not labeled as “curiously low” by Gresham et al. (2001) and was included in their mega-analysis. Consequently, we would like to provide reasons why we believe the exclusion of this meta-analysis was not justified.

First, Gresham et al. (2004) dismissed the Quinn et al. (1999) meta-analysis because the subjects were only “supposedly” (p. 38) EBD. However, Quinn et al. clearly
specified that subjects had to be identified as EBD using IDEA criteria; studies not meeting this provision were excluded. None of the five other meta-analyses included by Gresham et al. provided similar manifest EBD subject selection criteria. Thus, the literature base in the Gresham et al. mega-analysis includes five meta-analyses that do not specifically and exclusively focus on children with EBD, and one study excluded because it does.

Second, Gresham et al. (2004) questioned the EBD status of subjects from studies using single case designs. We believe that it is quite possible to identify single (i.e., individual) subjects as EBD using IDEA criteria. In many instances, subjects included in single-subject designs are those most in need of SST. Perhaps, if SST is to be an effective intervention, then working with these difficult subjects may be the best place to start.

Third, Gresham et al. (2004) excluded the Quinn et al. (1999) meta-analysis because many ESs were associated with dependent measures that included assessments for academic achievement and personality measures, not SST directly. But such outcome assessments can be found in the other meta-analyses (e.g., locus of control scales [Schneider & Byrne, 1985]; measures of academic achievement [Schneider, 1992]) which Gresham et al. included.

**Conclusion**

The efficacy of SST continues to be difficult to establish even though it is an increasingly popular intervention. The research has generally provided equivocal findings which, when synthesized with techniques like meta-analysis, tend to confirm that SST produces only modest effects. The expanding research base has permitted a number of meta-analyses to be conducted which has led to combining meta-analytic findings
themselves (i.e., mega-analysis). Although potentially providing significant insight, we believe that the mega-analyses provided by Gresham and colleagues do not effectively provide closure to questions about the efficacy of SST.

The failure of Gresham and colleagues to provide a rigorous and systematic methodology for their mega-analytic efforts, leaves many questions about the efficacy of SST unanswered, and precludes viewing SST as an increasingly effective intervention. In essence, efforts to enhance social competence have demonstrated only modest success and whether or not SST should be included in intervention programs because of increased effectiveness remains open to question.
Table 1

*Selection of SST Meta-analyses across Gresham and colleagues*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kavale, et al. (1997)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denham &amp; Almeida (1987)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forness &amp; Kavale, (1996)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathur, et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ang &amp; Hughes (2002)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Losel &amp; Beelmann (2003)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Schneider &amp; Byrne (1985)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quinn, et al. (1999)</td>
<td></td>
</tr>
</tbody>
</table>
References


