SYNTHESIS FINDINGS

Table 4 presents the meta-analytic results for the eight between groups quantitative studies in this review. One effect size (Hedges g) and accompanying weight (inverse variance method) were computed for each study. A weighted mean effect size was computed for the eight between groups studies, along with a *z* statistic to test for statistical significance and a *Q* statistic to evaluate for homogeneity of effect size estimates.

As shown in Table 4, the mean effect size for these studies was 0.55 and was statistically significant ($z = 5.6$; $p < .001$). This mean effect size would be considered a moderately large effect size according to Cohen's (1988) rubric. The Q-statistic for this group of eight studies, estimating the homogeneity of effect size estimates was not statistically significant ($Q = 9.17$; $df = 7$; $p = .24$) suggesting homogeneity of effect size estimates. Interestingly, every one of these between groups studies involved randomized or quasi- (or cluster) randomized participant assignment procedures. It is worth noting in Table 4 that three of the eight studies had confidence intervals associated with their effect sizes that included zero, an indication that those particular effect size estimates were not statistically significant in those three studies. The confidence interval associated with the average effect size, however, had a lower bound that was well above zero (.36) and a range that equaled the smallest range from among all of the confidence intervals in all eight studies.

Effect sizes were calculated for the single-participant and one group pretest-posttest design studies as well. With the single-participant designs, the effect sizes ranged quite broadly from 1.65 to 12.08. The average effect size was estimated to be 11.16 for these four studies. However the *Q*-statistic estimating homogeneity of effect size estimates was significant, reducing the confidence in the generalizability of this average effect size estimate. For example,

eliminating just the Presley and Hughes (2000) study reduced the average effect size to 2.79. However, in three of the four effect sizes, the lower limit of the confidence intervals associated with that effect size was negative, yet the average effect size was considered statistically significant. This anomaly undoubtedly has more to do with the instability of the effect size estimates (state of the field in effect size measurement for single participant designs) than as a valid representation of the effects of cognitive-behavioral interventions in single-participant study contexts.

Similarly, the fact that there were only two single group pretest-posttest design studies, as well as the relative weakness of these studies from an internal validity perspective severely limits the generalizability of the of the average effect size calculated from these two studies. It was positive, however, and was estimated to be 0.25.

Sensitivity Analyses

The single-participant and one group pretest-posttest design studies do not lend themselves to sensitivity analyses due to the extreme instability of the effect size estimates in the single-participant studies, and the limited number of one-group pretest-posttest design studies. With respect to participant characteristics the between groups studies were quite homogeneous on most characteristics such as gender, disability, and age range, again limiting the capacity for subgroup analyses. There was the opportunity to conduct a sensitivity analysis on the subgroup of three studies that employed quasi-randomized designs. These three studies employing quasi-randomization were the Barkley et al, (2001) study in which used an alternating assignment process described as follows:

> The first group of 12-14 sequentially referred families meeting eligibility criteria were enrolled in PSCT alone. As this treatment phase neared completion, the next

> wave of sequentially eligible families was then assigned to combined PSCT/BMP and so on in alternating waves. This method of assigning families to treatment was thus quasi-random (p. 927).

The other two quasi-randomized studies (Etscheidt, 1991; Robinson et al, 2002) used cluster randomization of intact classes to treatment conditions.

A sensitivity analysis conducted by analyzing the effect sizes of these three quasi-randomized studies and comparing the results of this analysis with an effect size analysis of the five randomized studies, yielded virtually the same mean effect size (0.53 and 0.58 respectively). The confidence intervals around the effect size estimates across these two analyses were very similar as well.

Rival Explanations

Although it is certainly possible that a number of rival explanations, dispersed across the 15 studies, could account for the effects found in this review it seems unlikely. First, the consistency of the results across all three design types (despite the instability of the effect size estimates in the single participant studies) reduces the probability of systematic or random error accounting for these findings. Second, and looking just at the strongest of the studies in this review – the between groups studies – the quality of the randomization in the designs and consistency of the magnitude, direction, and confidence intervals for those effect sizes tend to make rival explanations less defensible. Every one of these eight studies used some form of randomization – the single most important design characteristics that helps to rule out rival explanations.

Nonetheless, we feel compelled to point out several potential sources of bias that may make the results of this review less “bullet-proof” so-to-speak. The first is publication bias. All

of our studies were published in refereed journals. It is entirely conceivable that an unknown body of unpublished "file drawer" research with results counter to ours exists but was not published and hence was not accessible to us in our review. This is not a trivial issue in the conduct of systematic reviews like this one and must be considered as a potential source of downward pressure on effect size estimates that are found to be high and positive.

Second, despite the fact that we are confident in the technical validity with which we included (and excluded) single-participant studies on research design standards of internal and external validity, we are not comfortable with the interpretability of the effect size estimates derived from those studies that we did include. We recognize, too, that other meta-analysts in this fledgling area of research synthesis with single-participant studies may have used equally defensible alternative design standards for inclusion of studies, as well as other metrics for calculating effect sizes from those studies, and a different set of effect size estimates may have emerged.

Finally, from among the eight studies that used between groups designs, two (Etscheidt, 1991; Robinson et al, 2001) used cluster randomized assignment of classrooms to treatments. The appropriate statistical analysis for these types of designs is at the unit of assignment level (classrooms, rather than students). Nonetheless, both these studies reported analyses at the student level, diminishing somewhat the confidence that can be ascribed to the internal validity controls associated with the randomized nature of these designs.