DOCUMENT RESUME

ED 358 120 TM 019 893

AUTHOR Witt, Elizabeth A.
TITLE Meta-Analysis and the Effects of Coaching for Aptitude Tests.
PUB DATE Apr 93
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Aptitude Tests; *College Entrance Examinations; Comparative Analysis; Elementary Secondary Education; Higher Education; Intelligence Tests; Literature Reviews; Mathematical Models; *Meta Analysis; Pretests Posttests; Research Design; *Sampling; Scores; *Test Coaching; Testing Problems
IDENTIFIERS *High Stakes Tests

ABSTRACT

Two previous meta-analyses of the test-coaching literature have yielded conclusions that differ with respect to aptitude test scores. This study attempts to resolve differences in the conclusions of these two analyses by J. A. Kulik, R. L. Bangert-Drowns, and C. L. C. Kulik (1984) and K. Pearlman (1984) by using the methods of F. L. Schmidt and J. E. Hunter, which involve removal of variation due to sampling error, but modifying the basic procedures slightly to use formulas appropriate to the design of each study. Results suggest that small score gains might be expected from coaching and that high-stakes admissions tests are somewhat less amenable to coaching effects than are general intelligence tests. Within each of these groups, greater score gains were found when a pretest was administered in connection with the coaching program. The importance of both adjusting for sampling error (as did Hunter and Schmidt) and using design-appropriate variance formulas (as did Glass) is affirmed. Seven tables present study data. Contains 43 references. (SLD)

META-ANALYSIS AND THE EFFECTS OF COACHING FOR APTITUDE TESTS

Elizabeth A. Witt

The University of Iowa

April, 1992

Paper presented at the Annual Meeting
of the American Educational Research Association,
Atlanta, Georgia, April 12-16, 1993.

The author would like to thank Ken Pearlman of AT&T and Frank Schmidt and Robert Forsyth of the University of Iowa for their comments on an earlier version of this paper.

ABSTRACT

Two previous meta-analyses of the test-coaching literature have yielded conclusions that differ. Using Glassian meta-analytic techniques, Kulik, Bangert-Drowns, and Kulik (1984) determined that coaching programs do have a positive effect on aptitude test scores, but the effect is smaller for the SAT than for other aptitude tests. Pearlman (1984) employed the techniques of Hunter, Schmidt, and Jackson (1982), which involve removing variation due to sampling error; he found no appreciable difference between SAT and non-SAT coaching studies. Kulik and Kulik (1986), however, charged that Pearlman neglected to consider sample design when calculating variance due to sampling error.

This study attempts to resolve differences in the conclusions of these two previous meta-analyses by using the methods of Schmidt and Hunter, which involve removal of variation due to sampling error, but modifying the basic procedures slightly to employ formulas appropriate to the design of each study. Results suggest that small score gains might be expected from coaching and that high-stakes admissions tests are somewhat less amenable to coaching effects than are general intelligence tests. Within each of these groups, greater score gains were found when a pre-test was administered in connection with the coaching program. The importance of both adjusting for sampling error (ala Hunter and Schmidt) and employing design-appropriate variance formulas (ala Glass) is affirmed.

Aptitude test coaching has been a major topic of interest in the educational community over the past few decades. Every year new articles evaluating the effectiveness of coaching programs are added to the literature. While many report positive effects, results are mixed, and statistically significant effects are sometimes deemed too small to be of practical significance. Reviewers attempting to explain different findings often fail to reach similar conclusions, even when only considering studies involving a single aptitude test.

Much of the existing research has focused on the Scholastic Aptitude Test (SAT) (see Becker's [1990] recent meta-analysis). The SAT, however, is only one of a number of high-stakes tests for which coaching programs are widely offered. Consider, for example, the Graduate Record Examination, the Law School Admissions Test, and the Medical College Admissions Test. It seems reasonable to assume that the effects of coaching might be similar for tests like these; coaching techniques that work for one are likely to be effective for others as well. If this is not the case, it would be informative to discover which tests seem most (and least) susceptible to coaching effects and to consider the characteristics on which the groups of tests differ. A thorough meta-analysis, then, should include as much of the test-coaching literature as possible.

Kulik, Bangert-Drowns, and Kulik (1984) effectually searched the literature and produced a list of published and unpublished test-coaching studies conducted prior to 1982. This set of studies, which involves a variety of aptitude tests, was first

subjected to a Glassian meta-analysis by Kulik, Bangert-Drowns, and Kulik (1984), then reanalyzed with the methods of Hunter, Schmidt, and Jackson (1982) by Pearlman in 1984. This paper will attempt to resolve differences in the findings of these two meta-analyses.

The original study conducted by Kulik et al. (1984) examined all available aptitude test coaching studies whose design included a control group. Using Glassian techniques, they calculated a mean effect size and considered the relationship between effect size and several potential moderators; they did not weight studies according to sample size nor attempt to remove the effects of artifacts from the distribution of effect sizes. They concluded that coaching programs, in general, do have a positive effect on aptitude test scores, but the effect is much smaller for the SAT (mean effect size = .15 standard deviation) than for other aptitude tests (mean effect size = .43 standard deviation). Variation among the results of the non-SAT studies was explained in part by the inclusion or exclusion of a pre-test; other potential moderators were examined, but the researchers concluded that the variation among results of coaching studies was "impossible to explain fully" (Kulik et al., 1984, p. 187).

Pearlman (1984), however, suggested that much of the variation observed by Kulik et al. (1984) may be simply the result of sampling error. Using the meta-analytic procedures of Schmidt and Hunter (Hunter et al., 1982) to remove the effects of sam-

pling error from the mean and variance of the distribution of effect sizes, Pearlman reanalyzed the set of studies and found that 40% of the observed variance could be explained by sampling error alone. When mean effect sizes were calculated with sample-size weighting, no appreciable difference remained between SAT and non-SAT studies or between studies employing a pre-test and those using only a post-test. However, effect size was related to year of publication, with studies published prior to 1940 showing much greater effect sizes than those published since 1952.

Kulik and Kulik (1986), however, charged that Pearlman's (1984) results were inaccurate, claiming that he used the wrong formula in computing the variance due to sampling error for those studies employing something other than a post-test only, independent-groups design. They present alternate formulas, borrowed or derived from Hedges and Olkin (1985) or Glass et al. (Glass, McGaw, & Smith, 1981), for use with different designs. For studies involving the SAT, Pearlman found that 51% of the variance in observed effect sizes could be explained by sampling error. According to Kulik and Kulik, when the appropriate formulas are applied, sampling error accounts for only 12% of the variance in the SAT studies.

The same set of studies (SAT and non-SAT) is here reanalyzed by methods similar to those used by Pearlman (1984), but sampling error variance is computed by formulas appropriate to the experimental design of each study, as suggested by Kulik and Kulik

(1986). This paper also takes the analysis one step further than Pearlman's by accounting for the effects of measurement error (unreliability) on the distribution of observed effect sizes.[1]

This study was conceived as an exercise in meta-analysis, its main concern being to consider the importance of refinements in meta-analytic techniques. Researchers using the methods of Hunter et al. (1982), for example, often neglect to consider study design in computing the amount of variance due to sampling error and other artifacts, while researchers using Glassian procedures typically do not attempt to remove the effects of sampling error on the distribution of effect sizes before drawing their conclusions. Can such oversights significantly affect results?

A secondary, yet important concern of this study was to summarize the results of the literature on test-coaching. The original intention was to obtain all studies conducted since 1982 and include them with the earlier set in a comprehensive meta-analysis. Unfortunately, it soon became clear that, for this body of literature, the time and labor required not only to collect studies from various obscure sources but also to extract the necessary information for computing effect sizes would be prohibitive. Therefore, this paper, at least in its current edition, will not consider *all* test-coaching studies, but will concentrate on those conducted prior to 1982.

---

[1]Correcting for additional artifacts was considered impractical and/or of little importance.

PROCEDURE

The original search by Kulik et al. (1984) yielded 35 papers containing 38 studies. For the current study, these 35 papers--a mixture of dissertations, published articles, and technical reports--were obtained, and effect sizes were calculated afresh. The sample-size weighted mean effect size was computed. The studies were divided into groups according to their experimental design, and the sampling error variance was calculated separately for each group. The sample-size weighted average of the group error variances was used as the sampling error variance for the entire set of studies. Corrections were made for measurement error and for differences in reliability of the various aptitude tests using the procedures described by Hunter and Schmidt (1990; see pp. 311 ff.). Potential moderators were investigated by grouping the studies according to the characteristic of interest and conducting minor meta-analyses on each group.

One of the original 38 studies was omitted from this analysis. Although Alderman and Powers' (1980) main finding was an effect size of .08, this figure was apparently obtained by means of regression and Bayesian analyses; the error variance of this statistic could not be easily estimated and incorporated into the meta-analysis. Furthermore, the .08 effect size was apparently not independent of the data from which the second effect size (.12) was obtained. Therefore the .08 effect size was excluded, and only one study from the Alderman and Powers paper was counted.

The 38 studies are listed in Table 1 along with their respective designs, sample sizes, and effect sizes. Total sample sizes are shown for each study followed by a breakdown (in parentheses) into experimental and control group sample sizes. The overall sample size differs from that reported by Pearlman (1984). Pearlman obtained sample sizes from James Kulik; his figures probably reflect the total number of subjects involved in each study. The sample sizes shown in Table 1 were obtained directly from the original papers and include only those subjects supplying the data from which effect sizes were calculated.

In a few cases, the exact figures for experimental and control group sample sizes could not be determined but had to be estimated. Most studies gave sufficient information to make a reasonable estimate. The "worst case" of such estimates involved the two studies by the Federal Trade Commission (1979), where the total experimental group consisted of 1738 persons, and the total control group was comprised of 1003 subjects, but it was not clear how many people took part in each study. In this case each study was estimated to have used half of the subjects.

All studies involved both a control group and an experimental group. A few studies employed only a post-test; the majority used both a pre-test and a post-test. Effect sizes computed from the latter generally have a smaller standard error (as long as the correlation between pre-test and post-test is greater than .50). Note that the categorization of study design in Table 1 is merely dichotomous. Three studies actually used a

pre-post design with *matched* groups. Variance due to sampling error should be even further reduced by matching. Unfortunately, no formula for error variance could be located for this kind of design; instead, the formula for randomly assigned groups with a pre-post design served as a reasonable approximation. These three studies (nos. 5, 13, and 37) did not carry a lot of weight in the meta-analysis (total N=246). There were also several studies using analysis of covariance (ANCOVA) with pre-test scores and/or other variables as covariates. Kulik and Kulik (1986) provide yet another formula for computing the sampling error variance for effect sizes computed from this type of analysis. However, this formula is based on a formula for calculating effect size which, according to Glass et al. (1981), requires an additional term containing the regression coefficient. Without the final term, the obtained mean difference remains in the residual score metric, and the effect size cannot be directly compared with others in this meta-analysis. The regression coefficient was reported in almost none of these studies. Because of this and other difficulties in incorporating ANCOVAs into the meta-analysis, these studies were treated as ordinary pre-post designs; in most cases sufficient information was available for calculating effect sizes by the usual pre-post formula. Sampling error for these studies was then determined as for the pre-post studies.

Effect sizes were recomputed from the original studies and compared to those obtained by Kulik et al. (1984). This was

done, in part, to verify the design type and sample size used in computing each effect size. Following the procedures described by Kulik et al., pretest standard deviations were used in computing effect sizes wherever possible, while the control group standard deviation was used when there was no pretest. For pre-post studies Kulik and Kulik apparently computed separate effect sizes (standardized mean gains) for experimental and control groups, using each group's pretest standard deviation in finding its effect size, then subtracted the control group effect size from that of the experimental group to obtain the final effect size. Conversely, this analysis employed a pooled standard deviation as the denominator for both experimental and control groups' mean gain scores. (Pretest standard deviations of experimental and control groups were pooled.) The results are essentially the same. Effect sizes obtained by pooling are shown in the last column of Table 1. These are the effect sizes on which the current meta-analysis was performed. They are nearly identical to the values obtained by Kulik et al. (shown in the previous column), seldom differing by more than .01 and never by more than .03.

In a few instances it was unclear how Kulik et al. (1984) arrived at the figure they did. (For example, some articles reported neither standard deviations nor a t or F statistic.) In such cases the effect sizes reported by Kulik et al. were accepted without recomputation under the assumption that the research-

ers were able to obtain the necessary information from other sources.

Effect sizes for studies using only a post-test were obtained by the following formula:

$$d = (M_{E2} - M_{C2})/S_C$$

where

d = effect size

M_{E2} = mean of experimental group on post-test

M_{C2} = mean of control group on post-test

S_C = standard deviation of control group on post-test.

For studies employing both a pre-test and a post-test, effect sizes were found by:

$$d = [(M_{E2} - M_{E1}) - (M_{C2} - M_{C1})]/S_p$$

where d, M_{E2}, and M_{C2} are as described above and

M_{E1} = mean of experimental group on pre-test

M_{C1} = mean of control group on pre-test

S_p = pooled within-group standard deviation on pre-test.

If the appropriate means and standard deviations were not given, effect sizes for both types of study design were calculated from the t or F statistic as described by Glass et al. (1981). Sample-size weighted averaging yielded the mean effect size:

$$\bar{d} = \Sigma N_i d_i / \Sigma N_i .$$

The observed variance of effect sizes was computed by:

$$S_d^2 = \Sigma[N_i(d_i - \bar{d})^2]/\Sigma N_i .$$

Variance due to sampling error was computed separately for the two groups of studies, post-test only and pre-post designs.

Large sample formulas specifying experimental and control group N's were used for both groups since most studies involved over 100 subjects with different numbers assigned to experimental and control groups. For post-test only studies:

$$\text{post } S_e^2 = (1/\bar{N}_E + 1/\bar{N}_C) + \bar{d}^2/2\bar{N}_i$$

where

$\bar{N}_E$ = mean number of persons in experimental group per study

$\bar{N}_C$ = mean number of subjects in control group per study

$\bar{N}_i$ = mean number of total subjects per study.

This formula is equivalent to Hunter and Schmidt's large sample formula, $(4/\bar{N}_i)(1 + \bar{d}^2/8)$, when experimental and control group sample sizes are equal. (It should be noted that both formulas estimate sampling error variance based on the use of *pooled* estimates of the within-group standard deviations in computing the effect sizes. This study followed Kulik et al. [1984] in using control group standard deviations when given. Therefore these formulas underestimate slightly the actual amount of variance that can be explained by sampling error in the effect sizes obtained from this set of post-test only studies. Thus the use of these formulas is conservative in this context.)

Error variance for pre-post studies was computed thus:

$$\text{pre-post } S_e^2 = 2(1 - \bar{r}_{12})(1/\bar{N}_E + 1/\bar{N}_C) + \bar{d}^2/2\bar{N}_i$$

where $\bar{r}_{12}$ is the mean correlation between pre-test and post-test scores. The pre-post correlation was reported in only a few studies; however, reasonable estimates were available in the form of test-retest reliabilities (see Glass et al., p. 118) for most

studies.[2] These were averaged across tests to find $\bar{r}_{12}$. Finally the error variance for the entire set of studies was found by a weighted average of the group error variances:

$$S_e^2 = (\Sigma N_{group} S_{e\,group}^2)/\Sigma N_{group}.$$

Variance not accounted for by sampling error was then obtained by subtracting S_e^2 from S_d^2, and the proportion of variance accounted for by sampling error was observed in the ratio S_e^2/S_d^2.

The procedures described thus far constitute a bare-bones meta-analysis; the only artifact corrected for was sampling error. The next step was to correct for the effects of unreliability in the aptitude tests. Reliabilities were obtained for all but five tests by perusing test manuals and testing reference books. It would have been possible to correct each effect size individually, substituting the mean reliability for the five missing values. However, concerns for expediency and computational simplicity dictated a less tedious approach (at least for this edition of the paper). Moreover, the reliabilities of these tests are high (indicating that correction of effect sizes would nct be expected to make a great difference in the outcome) and similar to one another (so that the mean value is seldom far off from the actual reliability). Nearly all reliabilities are in

---

[2]Test-retest reliabilities would overestimate pre-post correlations in the case of a subject-by-treatment interaction; the amount of variance attributable to sampling error would then be underestimated--a conservative move in this context. Many of these coaching studies, however, either assumed no interaction or tested for interaction and found none. Therefore, any inaccuracy due to the use of test-retest reliabilities in the place of pre-post correlations is likely to be minor.

the upper .80s and lower .90s, with the full range extending only from .83 to .97. A single overall correction of the mean effect size was deemed appropriate:

$$\bar{d}_t = \bar{d}/\mathrm{SQRT}(\bar{r}_{yy})$$

where $\mathrm{SQRT}(\bar{r}_{yy})$ denotes the square root of the mean reliability of the aptitude tests and $\bar{d}_t$ represents the mean true (disattenuated) effect size.

The variance of effect sizes was corrected for error of measurement using the distribution of the square roots of reliabilities as described by Hunter and Schmidt (1990; p. 312):

$$S_{dt}^2 = [(S_d^2 - S_e^2) - \bar{d}_t^2 S_a^2]/\bar{r}_{yy}$$

where S_{dt}^2 is the variance remaining after variance due to both sampling error and measurement error have been removed and S_a^2 is variance of the distribution of the square roots of the reliabilities. The latter term in this analysis was so small that, after rounding, it had no effect on the outcome; therefore,

$$S_{dt}^2 \cong (S_d^2 - S_e^2)/\bar{r}_{yy}.$$

Because correcting for the effects of measurement error on the distribution of effect sizes produced only very small changes, bare-bones meta-analyses were considered sufficient for investigating potential moderator variables. Studies were divided into groups according to the characteristic of interest, and each group was subjected to a meta-analysis as described above, correcting only for sampling error.

RESULTS

The results of the main meta-analyses are shown in Table 2. The best estimate of observed score gain resulting from aptitude test coaching is .22 standard deviation. Sampling error accounts for approximately 19% of the variance in observed effect sizes, leaving a considerable amount unexplained. It is highly unlikely that all of the coaching programs investigated by this set of studies produce effects of similar magnitude. The 90% credibility interval for the test coaching effect size includes zero, suggesting that some coaching programs may have little or no positive effect on aptitude test scores.

The best estimate of true score gain resulting from coaching is .23 standard deviation. The scant difference between this figure and the observed effect size is not surprising since the mean reliability of these aptitude tests is .916. Correcting for measurement error decreased the proportion of variance accounted for to 11%. (Essentially, the act of correcting added more variance than difference in reliabilities accounted for.) The direct contribution to variance made by differences in reliability among the various aptitude tests was a minuscule .000016 ($= .23^2 \times .0003$). This is one case where, because test reliabilities were high and not very variable, correcting for measurement error was hardly worth the labor involved. Therefore moderators were considered using bare-bones procedures alone.

Kulik et al. (1984) found differences between effect sizes depending on whether or not a pre-test was included in the study

design and on whether or not the dependent variable was the SAT. Pearlman (1984) concluded that the disparities between SAT and non-SAT studies and between post-test and pre-post studies were too small to warrant attention, given the degree of overlap in 95% confidence intervals[3] for each group's effect size. He did, however, find year of publication to be a significant moderator, with effect sizes from studies published prior to 1940 higher than those from studies published in 1952 or later. No other moderator variables manifested themselves in either of these previous meta-analyses.

The 37 studies are listed in order of effect size in Table 3, each accompanied by its year of publication, study design classification, and dependent variable (aptitude test). A glance at the column of tests reveals a clustering of most of the SAT studies in the lower half of the distribution. However, certain other tests with effect sizes below .40--namely the National Medical Board, the GRE-Q, and the MCAT--have something in common with the SAT. These tests are all used primarily for admission to higher education or to a profession; they are high-stakes tests. All are administered chiefly to young adults. All measure abilities which have developed primarily as a result of

[3]What Pearlman refers to as a confidence interval might more accurately be termed a "credibility interval." It is an interval constructed around the mean effect size based not on sampling error but on the amount of variance remaining once sampling error (and variance due to other artifacts) has been removed. The distribution of interest is one of differing population effect sizes, not one of differing sample effect sizes drawn from a common population. The latter distribution would be described in terms of a confidence interval. See Whitener (1990).

schooling. These high-stakes admissions tests are classified in Table 3 as "Type 1." The remaining tests tend to focus instead on general intelligence and reasoning abilities which may well have developed outside of an academic setting. They are more likely to be used for classification or diagnostic purposes than for admissions. For lack of a better descriptor, these tests are referred to as "general intelligence tests" and classified as "Type 2."

Minor meta-analyses were conducted on the Type 1 and Type 2 tests separately. The results are shown in Table 4. There is a notable difference in mean effect sizes: .18 for the high-stakes tests and .33 for the general intelligence tests. The standard deviation of effect sizes (after removal of sampling error) was reduced from .1682 in the total group to .1356 and .1646 for the Type 1 and Type 2 groups, respectively. In addition, a considerable portion of the variance among Type 2 effect sizes is now explained by sampling error alone. Thus there are indications that this classification of tests--high-stakes admissions tests versus measures of general intelligence--constitutes a genuine moderator, with Type 2 tests being somewhat more coachable than the high-stakes tests. At the same time, it should be kept in mind that there still remains a lot of unexplained variance, especially among Type 1 tests, and the 90% credibility intervals for the two groups overlap considerably. Thus, while there is some evidence of a moderator, its existence (and nature) should be accepted with caution.

The 37 studies were next grouped according to study design: post-test only versus pre- and post-tests. Table 5 shows the outcome of the group meta-analyses. Again, there is a notable difference in mean effect sizes; a larger average coaching effect is obtained in studies including a pre-test. However, the standard deviation of effect sizes actually increased for the pre-post group; in fact, the average (pooled) standard deviation for the two groups is slightly larger than the total group standard deviation of .1682.[4] This suggests that a true moderator has *not* been uncovered; if the two groups did in fact have different population means, the average within-group variance should be smaller than the variance for the total group. If a moderating variable is at work here, it must be something more complex than this classification of study design. It is interesting to note, however, that the 90% credibility interval for the pre-post studies does not include zero; we can be reasonably certain any coaching study including a pre-test will find a rise test scores. The magnitude of the gain may vary, but on the average our best estimate of the expected increase is a little more than a quarter of a standard deviation.

It seemed plausible that the variance within study-design groups may have been affected by the distribution of test type

---

[4]Group variances were weighted by number of studies in calculating the pooled standard deviation. Weighting instead by the total number of subjects in each group would not alter the conclusion: the average group variance is still larger than the total variance.

over the study-design groups; in other words, it may have been inflated by an interaction. Therefore post-test versus pre-post designs were examined within test type, with meta-analyses conducted separately on each subgroup. The results are shown in Table 6.

High-stakes tests show a modest difference in mean effect size between post-test and pre-post designs: .10 for studies using only a post-test, .20 for studies using both pre- and post-tests. Variation among post-test studies is reduced to zero once sampling error has been accounted for. The standard deviation of effect sizes from pre-post studies also decreased, though only slightly (.1327, compared to .1356 from Table 4). Hence study design appears to be a viable moderator among Type 1 tests. Coaching programs for high-stakes admissions tests might be expected to raise observed scores by about one-tenth of a standard deviation if no pre-test is used as part of the coaching. Programs with a pre-test *may* raise scores a little more (best estimate of .20 standard deviation); however, 97% of the variance among pre-post studies has not been accounted for, and the 90% credibility interval for this group is quite broad. For reasons as yet unclear, programs involving pre-tests in coaching for high-stakes tests have apparently worked better in some situations than in others.

The lower half of Table 6 shows the results for meta-analyses conducted by study-design group for Type 2 (general intelligence) tests. Here a large difference in mean effect sizes is

revealed. The mean for Type 2 post-test studies is the same as the mean for Type 1 post-test studies: .10. Pre-post designs, on the other hand, averaged a respectable .48, a figure most people would consider to be of practical significance. The standard deviation for the post-test group is reduced to .1584 (down from .1646 in Table 4), while the standard deviation of the pre-post group is unchanged from the total group figure for Type 2 tests. The difference in means coupled with the reduction in average variance indicates that study design is a moderator among Type 2 tests as well. Over one-third of the observed variance in each study-design group is explained by sampling error. The 90% credibility intervals do overlap, but to a much lesser extent than has occurred previously in this analysis. Perhaps more importantly, the lower tail of the credibility interval for pre-post studies rests well above the zero point. Whitener (1990) suggests that no further moderators are required to explain the remaining variance of effect sizes when the credibility interval does not approach zero. Thus, when the measurement of coaching effects for a general intelligence test includes the use of a pre-test, scores can be expected to rise, and the gain may be large enough to make a meaningful difference. The score gain to be expected is, on the average, about half a standard deviation.

Minor meta-analyses were *not* conducted on studies grouped according to publication year. All five of the pre-1940 studies are classified as Type 2 in Table 3. Their effect sizes are fairly evenly spread over the distribution of Type 2 studies.

Although there is a difference in the mean effect sizes of earlier and later Type 2 studies (.55 for pre-1940; .39 for post-1951), it appears to be due largely to the presence of a couple of heavily-weighted, low-lying outliers among the Type 2 tests from the later period (studies 32 and 12). If the lowest outlier (the only Type 2 study with a negative effect size) is ignored, the mean effect size for the remaining post-1951, Type 2 tests becomes .56, essentially equal to the mean for pre-1940 studies. Furthermore, if Type 2 studies were grouped according to early versus "modern" times and outliers were not removed, it seems unlikely that the standard deviation of the later group would be reduced enough to justify the identification of a moderator.[5]

Kulik et al. (1984) list several other moderators which have theoretical plausibility. The 35 papers from which the studies were obtained were perused (admittedly somewhat lightly) with these potential moderators in mind. These included, among other things, the content and duration of coaching, the source and sponsor of the program, the age and ability of the subjects, the type of aptitude test, and publication characteristics. No study characteristics appeared to be systematically related to effect size, even with studies grouped by test type and study design. Still, a fair amount of variance among effect sizes remains

---

[5]If the data analysis were to be handled by the computer, it may be worthwhile to perform minor meta-analyses according to year of publication. However, the calculations for this study were performed by hand (appropriate software is not yet available), and year of publication did not exhibit enough potential to justify the labor required for further investigation.

unexplained, especially among pre-post studies using high-stakes tests. A truly perfectionistic meta-analyst might be tempted to delve a little deeper. True perfectionists, however, rarely survive graduate school. Since the major points of disagreement between Pearlman (1984) and Kulik et al. (1984) have been satisfactorily resolved, this meta-analysis stops here.

DISCUSSION

To a degree, the results of this study support the conclusions of both previous meta-analyses. Because the majority of these coaching studies employed some type of pre-post design, the proportion of variance due to sampling error was considerably less than Pearlman (1984) had estimated; consequently, this study's conclusions regarding moderator variables are more in line with those of Kulik et al. (1984) than with Pearlman's. The magnitude of a coaching effect depends in part on the type of aptitude test; within test type, effect sizes differ depending on whether or not the study design included a pre-test. Mean effect sizes, however, are very similar to those found by Pearlman; they are generally lower than the means obtained by Kulik et al. Mean effect sizes are listed in Table 7. Comparisons with figures in the third column should be made cautiously. These figures were obtained from 37 of the original 38 studies, and sample sizes differed slightly from those used by Pearlman. In addition, the classification of test type in the final meta-analysis was not strictly SAT versus non-SAT; three additional studies with high-stakes tests were grouped with the SAT studies.

If the mean effect sizes shown in Tables 6 and 7 by study design within test type are accepted as the best estimates of gains to be expected as a result of coaching, coaching programs in studies using a pre-test appear to result in higher gains. This could be simply a reflection of the greater power of the pre-post study design. However, Kulik et al. (1984) suggest that a pre-test may be an effective component of test preparation. It is possible that the use of a pre-test as part of a coaching program could raise scores, perhaps by serving to reduce examinees' anxiety and minimize the mental energy they must invest in familiarizing themselves with the peculiarities of the particular test on a later administration. On the other hand, it may be that pre-tests themselves have little effect but are simply associated somehow with coaching programs employing better techniques. Researchers who design better studies may also design more effective coaching programs.

With or without a pre-test, however, coaching programs may differ somewhat in their effectiveness; if a student is considering enrolling in a coaching program for a high-stakes test, he or she would do well to ask for evidence of its effectiveness. Publishers of high-stakes tests generally agree that an observed score increase of .10 to .20 standard deviation is not enough to give coached students an unfair advantage and do not believe that such a gain is worth the fee often charged by test coachers. However, students whose ability level hovers just below the performance required to meet an admission cut-off may find that

coaching is well worth the cost, especially if their gain is enhanced beyond the mean because of high motivation or positive measurement error (simple luck).

Regarding the difference in effect sizes for Type 1 and Type 2 tests, the results suggest that if a student is applying to graduate school, for example, and is required to submit scores on both the Graduate Record Exam (GRE) and the Miller Analogies Test (MAT), he or she might do better by devoting more attention to coaching for the MAT than for the GRE (assuming the tests are weighted equally in the admissions process), since coaching appears to be more likely to produce a meaningful increase in score on an analogies test, especially when a pre-test is a part of the coaching program. In fact, anyone expecting to take a general intelligence test for high-stakes purposes would be wise to seek coaching (with a pre-test), as his or her score is likely to rise by at least a fifth of a standard deviation.

As K. Pearlman (personal communication, January 6, 1993) pointed out, nearly all the tests included in this set of studies measure some combination of the verbal, quantitative, and analytic abilities that contribute to general intellectual ability. It seems contrary to expectation that tests of general intelligence and aptitude should be more coachable than tests that measure abilities developed primarily as a result of formal education. Furthermore, because coaching programs for high-stakes admissions tests have existed in abundance for a long time, one would think there have been many opportunities to improve the coaching

procedures associated with these tests and capitalize on unique, coachable features of each test. One possible explanation is that high-stakes admissions tests, in comparison with Type 2 tests, are so well known that even examinees who receive no coaching are already familiar with their content, item types, and other features so that coaching adds relatively little new information to the knowledge most examinees already possess. Another possibility is that the group of tests here designated as Type 1 are less coachable simply because they are more likely to be coached. The higher the stakes, the more examinees are likely to seek coaching; and the more test coachers stand to profit, the more coaching programs are offered. Publishers of large-scale, widely used tests naturally do not want their instruments to be sensitive to short-term effects of coaching, and they are aware of the plethora of coaching programs that arise wherever their tests are administered. Perhaps they therefore put more effort into deliberately designing their tests to make it difficult for coachers to directly teach to the test.

The difference in the results of this study compared with the two previous meta-analyses confirms the importance of sample-size weighting in meta-analysis as well as the use of design-appropriate formulas for computing the sampling error in observed distributions of effect sizes. Disregarding either of these refinements may lead to a substantive difference in one's conclusions.

This study also suffers a number of limitations. Foremost of these is the fact that only studies published prior to 1982 were included in the analysis. The nature of coaching may have changed in the past decade; certainly one would expect that those who make their living coaching for standardized tests would be continually attempting to perfect their craft. Conversely, test publishers may be continually attempting to build tests that are less amenable to coaching. Before any definitive statements can be made about the effectiveness of test coaching today, studies from the past decade must be examined. The inclusion of more recent studies would also greatly increase the size of the sample; once the present set of studies is grouped by test type and by the inclusion or exclusion of a pre-test, the numbers of studies in some groups are very small. More confidence could be placed in the conclusions of this meta-analysis if a greater number of studies were included. The implementation of a comprehensive meta-analysis of *all* test-coaching studies conducted to date using the techniques outlined in this study is an obvious next step.

Another potential problem is the fact that a meta-analysis of any literature as diverse as test-coaching studies requires the researcher to make a host of difficult, arbitrary decisions. Among studies unearthed in the search process, some will inevitably be eliminated; by what criteria should studies be judged? Given a set of studies, what choice of standard deviation is most suitable for computing effect sizes that can be meaningfully

compared across studies? How should the combination of dependent data be handled? When studies report score differences for different groups and/or different exams and/or two or more analyses involving *parts of* the same groups, for example, what is the best method of computing an effect size that can be compared with the effect sizes found in other studies? What data (if any) should be ignored, and why? Should outliers be removed? If so, how does one define an outlier? If insufficient data are provided, under what circumstances is it reasonable to compute effect sizes using estimates of missing data obtained from other sources? Questions like these seldom have a single best answer, and the research on test coaching is perhaps more vulnerable to the effects of subjectivity in these areas than are other, less complex bodies of literature. Kulik et al. (1984) were reasonably thorough in describing the decision rules and procedures they followed for their search and analysis. Yet, even following their procedures, this researcher was not always inclined to agree with their decisions.

Finally, for this meta-analysis, the decision was made to treat studies employing a matched pre-post design or an analysis of covariance (ANCOVA) as if they were random pre-post designs. The variance attributable to sampling error is probably somewhat overestimated for the three studies involving matched groups. The effect on the outcome of the analysis is probably minimal, since these studies together did not carry much weight. Nevertheless, it would be preferable to derive the appropriate formula

and obtain a more accurate estimate of sampling error for matched pre-post designs. For studies using ANCOVA, sampling error was computed in a manner appropriate to the calculation of effect sizes, which were computed as straightforward standardized differences between mean gains. Presumably more precision could be gained by using the error variance for residual scores; unfortunately nearly all studies failed to report enough information to allow the calculation of this statistic.

In all likelihood, more information exists on any research topic than can be extracted from the literature for meta-analytic review because the necessary figures are not always reported. Perhaps as researchers and publishers become increasingly aware of the usefulness of meta-analysis, there will evolve a tendency to report results in terms of effect sizes which are compatible with others in the same subfield and to include all the information necessary for an accurate compilation of data via meta-analytic techniques. Meanwhile, meta-analysis remains a useful tool (if not as precise as we would like) for synthesizing the diverse conclusions of a body of research and at least approximating the distribution of effect sizes. At the same time, it is a tool that requires careful handling.

REFERENCES

Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-verbal scores. *American Educational Research Journal, 17,* 239-253.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60,* 373-417.

Bernal, E. M., Jr. (1971). Concept learning among Anglo, Black, and Mexican-American children using facilitation strategies and bilingual techniques. *Dissertation Abstracts International, 32,* 6180A. (University Microfilms No. 72-15,707)

Boger, J. H. (1952). An experimental study of perceptual training on group IQ test scores of elementary pupils in rural ungraded schools. *Journal of Educational Research, 46,* 43-52.

Casey, M. L., Davidson, H. P., & Horter, D. I. (1928). Three studies on the effect of training in similar and identical material upon Stanford-Binet test scores. *Twenty-seventh Yearbook of the National Society for the Study of Education, 1,* 431-439.

Evans, F. R. (1973). *The GRE-Q Coaching/Instruction Study.* Princeton, NJ: Graduate Record Examinations, Educational Testing Service. (ERIC Document Reproduction Service No. ED 163 088)

Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10,* 257-272.

Federal Trade Commission, Bureau of Consumer Protection. (1979). *Effects of coaching on standardized admission examinations: Revised statistical analyses of data gathered by Boston Regional Office of the Federal Trade Commission.* Washington, DC: Federal Trade Commission, Bureau of Consumer Protection. (NTIS No. PB-296 196)

Flynn, J. T., & Anderson, B. E. (1977). The effects of test item cue sensitivity on IQ and achievement test performance. *Educational Research Quarterly, 2*(2), 32-39.

Frankel, E. (1960). Effects of growth, practice, and coaching on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 38,* 713-719.

French, J. W. (1955). *The coachability of the SAT in public schools* (RB 55-26). Princeton, NJ: Educational Testing Service.

French, J. W., & Dear, R. E. (1959). Effect of coaching on an aptitude test. *Educational and Psychological Measurement, 19,* 319-330.

Gilmore, M. F. (1927). Coaching for intelligence tests. *Journal of Educational Psychology, 18,* 119-121.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Goldsmith, R. P. (1980). The effects of training in test taking skills and test anxiety management on Mexican American students' aptitude test performance. *Dissertation Abstracts International, 40,* 5790A. (University Microfilms No. 80-09863)

Greene, K. B. (1928). The influence of specialized training on tests of general intelligence. *Twenty-seventh Yearbook of the National Society for the Study of Education, 1,* 421-428.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Holloway, H. D. (1954). Effects of training on the SRA Primary Mental Abilities (Primary) and the WISC. *Child Development, 25,* 253-263.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Jefferson, J. L. (1975). The effects of anxiety on the achievement of black graduate students taking standardized achievement tests. *Dissertation Abstracts International, 35,* 5121A. (University Microfilms No. 75-3105)

Keefauver, L. W. (1977). The effects of a program of coaching on Scholastic Aptitude Test scores of high school seniors pretested as juniors. *Dissertation Abstracts International, 37,* 5063A. (University Microfilms No. 77-3651)

Keysor, R. E. (1977). The effect of test wiseness on professional school screening test scores. *Dissertation Abstracts International, 37,*(9-B) 4652. (University Microfilms No. 77-4834)

Kintisch, L. S. (1979). Classroom techniques for improving Scholastic Aptitude Test scores. *Journal of Reading, 22,* 416-419.

Klutch, M. I. (1976). The influence of test sophistication on standardized test scores. *Dissertation Abstracts International, 37,* 809A. (University Microfilms No. 76-19,058)

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95,* 179-188.

Kulik, J. A., & Kulik, C. L. C. (1986, April). *Operative and interpretable effect sizes in meta-analysis.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 276 759)

Lent, R. W., & Russell, R. K. (1978). Treatment of test anxiety by cue-controlled desensitization and study-skills training. *Journal of Counseling Psychology, 25,* 217-224.

Lewis, L. A., & Kuske, T. T. (1978). Commercial National Board review programs: A case study at the Medical College of Georgia. *Journal of the American Medical Association, 240,* 754-755.

Melametsa, L. (1965). The influence of training on the level of test performance and the factor structure of intelligence tests. *Scandinavian Journal of Psychology, 6,* 19-25.

Merriman, C. (1927). Coaching for mental tests. *Educational Administration and Supervision, 13,* 59-64.

Moore, J. C. (1971). Test-wiseness and analogy test performance. *Measurement and Evaluation in Guidance, 3,* 198-202.

Oakland, T. (1972). The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children. *Journal of School Psychology, 10,* 355-360.

Petty, N. E., & Harrell, E. H. (1977). Effect of programmed instruction related to motivation, anxiety, and test wiseness on group IQ test performance. *Journal of Educational Psychology, 69,* 630-635.

Pearlman, K. (1984, August). Validity generalization: Methodological and substantive implications for meta-analytic research. In H. Wing (Chair), *Meta-analysis: Procedures, practices, and pitfalls.* Symposium conducted at the meeting of the American Psychological Association, Toronto, Ontario, Canada.

Rayford, O. L. (1973). An experimental study of the effects of three modes of test orientation on scholastic aptitude and achievement scores. *Dissertation Abstracts International, 33,* 6099A. (University Microfilms No. 73-12,646)

Roberts, S. O., & Oppenheim, D. B. (1966). *The effect of special instruction upon test performance of high school students in Tennessee.* Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service, No. ED 053 158)

Rutan, P. C. (1979). Test sophistication training: A program level intervention for the school psychologist. *Dissertation Abstracts International, 40,* 171A. (University Microfilms No. 79-14,135)

Trainor, J. C. (1939). Experimental results of training in general semantics upon intelligence test scores. *Papers from the First American Congress on General Semantics--Ellensberg, Washington.* New York, NY: Arrow Editions.

Whitely, S. E., & Dawis, R. V. (1974). Effects of cognitive intervention on latent ability measured from analogy items. *Journal of Educational Psychology, 66,* 710-717.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315-321.

Whitla, D. K. (1962). Effect of tutoring on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 41,* 32-37.

Wiseman, S., & Wrigley, J. (1953). The comparative effects of coaching and practice on the results of verbal intelligence tests. *British Journal of Psychology, 44,* 83-94.

TABLE 1

Study	Design	N (N_E;N_C)	Kulik et al. Effect Size	Effect Size
1. Alderman & Powers (1980)	PRE-POST	559 (239;320)	.08	OMIT
2. Alderman & Powers (1980)	POST	559 (239;320)	.12	.12
3. Bernal (1971)	POST	192 (96;96)	.40	.41
4. Boger (1952)	PRE-POST	104 (54;50)	.45	.45
5. Casey, Davidson, & Horter (1928)	PRE-POST	52 (26;26)	.74	.74
6. Dear (cited in French & Dear, 1959)	PRE-POST	187 (71;116)	.26	.26
7. Dyer (cited in French & Dear, 1959)	PRE-POST	418 (225;193)	.10	.10
8. Evans (1977)	PRE-POST	254 (88;166)	.26	.26
9. Evans & Pike (1973)	PRE-POST	502 (337;165)	.52	.52
10. Federal Trade Commission (1979)	PRE-POST	1371 (869;502)	.31	.31
11. Federal Trade Commission (1979)	PRE-POST	1370 (869;501)	.07	.07
12. Flynn & Anderson (1977)	POST	181 (90;91)	.02	.02
13. Frankel (1960)	PRE-POST	90 (45;45)	.10	.10
14. French (1955)	PRE-POST	429 (216;213)	.17	.17
15. French (1955)	POST	381 (169;212)	.05	.05
16. Gilmore (1927)	PRE-POST	64 (32;32)	.71	.71
17. Goldsmith (1980)	PRE-POST	114 (50;64)	.66	.67
18. Greene (1928)	PRE-POST	155 (102;51)	.57	.55
19. Holloway (1954)	PRE-POST	107 (53;54)	.53	.52
20. Jefferson (1975)	PRE-POST	50 (25;25)	.70	.69
21. Keefauver (1977)	PRE-POST	52 (16;36)	-.01	.00
22. Keysor (1977)	POST	166 (54;112)	.18	.18
23. Kintisch (1979)	PRE-POST	76 (38;38)	.14	.14
24. Klutch (1976)	PRE-POST	80 (29;51)	.43	.46
25. Lent & Russell (1978)	PRE-POST	57 (31;26)	.44	.44
26. Lewis & Kuske (1978)	PRE-POST	133 (33;100)	-.06	-.06
27. Melametsa (1965)	PRE-POST	130 (69;61)	.84	.84
28. Merriman (1927)	PRE-POST	105 (50;55)	.40	.40
29. Moore (1971)	POST	38 (19;19)	.78	.75
30. Oakland (1972)	PRE-POST	61 (36;25)	.46	.46
31. Petty & Harrell (1977)	PRE-POST	47 (24;23)	.23	.23
32. Rayford (1973)	POST	638 (425;213)	-.01	-.01
33. Roberts & Oppenheim (1966)	PRE-POST	688 (342;346)	.12	.13
34. Rutan (1979)	PRE-POST	47 (22;25)	.57	.57
35. Trainor (1939)	PRE-POST	30 (15;15)	.45	.45
36. Whitely & Dawis (1974)	PRE-POST	65 (34;31)	.43	.43
37. Whitla (1962)	PRE-POST	104 (52;52)	.03	.03
38. Wiseman & Wrigley (1953)	PRE-POST	269 (129;140)	.13	.13
	TOTAL N:	9366 (5074;4292)		

TABLE 2

META-ANALYSES FOR THE ENTIRE SET OF STUDIES

Bare-Bones (correcting for sampling error only)	Full Analysis (correcting for sampling and measurement error)
$\bar{d} = .22$	$\bar{d}_t = .23$
$S_d^2 = .0348$	$S_d^2 = .0348$ (from bare-bones)
$S_e^2 = .0065$	$S_a^2 = .0003$
$S^2 = S_d^2 - S_e^2 = .0283$	$S_{dt}^2 = .0309$
$S = .1682$	$S_{dt} = .1758$
$S_e^2/S_d^2 = .187$ or 19%	$(S_d^2 - S_{dt}^2)/S_d^2 = .112$ or 11%
90% credibility interval for d: (-.06, .50)	90% credibility interval for d: (-.06, .52)

$\bar{d}$ = mean effect size

$\bar{d}_t$ = mean effect size corrected for measurement error

S_d^2 = total variance of observed effect sizes

S_e^2 = variance due to sampling error

S^2 = variance remaining after sampling error is accounted for

S = square root of S^2 (standard deviation)

S_a^2 = variance of attenuation factors (square roots of reliabilities)

S_{dt}^2 = variance remaining after correcting for sampling error and measurement error ($S_{dt}^2 = [(S_d^2 - S_e^2) - d_t^2 S_a^2]/r_{yy}$)

S_{dt} = square root of S_{dt}^2 (standard deviation)

TABLE 3

STUDY CHARACTERISTICS

Study	Year publ.	N	d	Design	Test(s)	Test type
26	1978	381	-.06	PRE-POST	National Medical Board	1
32	1973	638	-.01	POST	Lorge-Thorndike	2
21	1977	52	.00	PRE-POST	SAT	1
12	1977	181	.02	POST	Thurstone Test of Mental Alertness	2
37	1962	104	.03	PRE-POST	SAT	1
15	1955	381	.05	POST	SAT-V	1
11	1979	1370	.07	PRE-POST	SAT	1
7	1959	418	.10	PRE-POST	SAT-M	1
13	1960	90	.10	PRE-POST	SAT	1
2	1980	559	.12	POST	SAT-V	1
33	1966	688	.13	PRE-POST	PSAT	1
38	1953	269	.13	PRE-POST	Moray House Intell. Test	2
23	1979	76	.14	PRE-POST	SAT-V	1
14	1955	429	.17	PRE-POST	SAT	1
22	1977	166	.18	POST	MCAT	1
31	1977	47	.23	PRE-POST	Otis-Lennon	2
6	1959	187	.26	PRE-POST	SAT-M	1
8	1977	254	.26	PRE-POST	GRE-Q	1
10	1979	1371	.31	PRE-POST	SAT	1
28	*1927	105	.40	PRE-POST	Thorndike Intell. Exam	2
3	1971	192	.41	POST	SRA Primary Mental Abil.	2
36	1974	65	.43	PRE-POST	Teacher-made (analogies)	2
25	1978	57	.44	PRE-POST	Teacher-made (analogies & digit symbol)	2
4	1952	104	.45	PRE-POST	Otis Quick-Scoring & Cal. Test of Mental Maturity	2
35	*1939	30	.45	PRE-POST	Detroit Intell. Test	2
24	1976	80	.46	PRE-POST	DAT	2
30	1972	61	.46	PRE-POST	Metropol. Readiness Test	2
9	1973	502	.52	PRE-POST	SAT-M	1
19	1954	107	.52	PRE-POST	SRA Primary Mental Abil.	2
18	*1928	155	.55	PRE-POST	Stanford-Binet	2
34	1979	47	.57	PRE-POST	DAT	2
17	1980	114	.67	PRE-POST	DAT-V	2
20	1975	50	.69	PRE-POST	Otis-Lennon	2
16	*1927	64	.71	PRE-POST	Otis Group Intell. Scale	2
5	*1928	52	.74	PRE-POST	Teacher-made (Stanford-Binet plus other items)	2
29	1971	38	.75	POST	Teacher-made (analogies)	2
27	1965	130	.84	PRE-POST	Teacher-made (number patterns)	2

*Published prior to 1940

TABLE 4

META-ANALYSES OF GROUPED STUDIES:
HIGH-STAKES ADMISSION VS. GENERAL INTELLIGENCE TESTS

Type 1: High-stakes tests	Type 2: General intellig. tests
No. of studies = 16	No. of studies = 21
Total N = 6780	Total N = 2586
$\bar{d} = .18$	$\bar{d} = .33$
$S_d^2 = .0206$	$S_d^2 = .0830$
$S_e^2 = .0022$	$S_e^2 = .0559$
$S^2 = S_d^2 - S_e^2 = .0184$	$S^2 = S_d^2 - S_e^2 = .0271$
$S = .1356$	$S = .1646$
$S_e^2/S_d^2 = .107$ or 11%	$S_e^2/S_d^2 = .6735$ or 67%
90% credibility interval for d: (-.04, .40)	90% credibility interval for d: (.06, .60)

(Notation is explained in Table 2.)

TABLE 5

META-ANALYSES OF GROUPED STUDIES:
POST-TEST ONLY VS. PRE-POST DESIGNS

Studies with post-test only	Studies with pre- and post-tests
No. of studies = 7	No. of studies = 30
Total N = 2155	Total N = 7211
$\bar{d} = .10$	$\bar{d} = .26$
$S_d^2 = .0212$	$S_d^2 = .0389$
$S_e^2 = .0130$	$S_e^2 = .0045$
$S^2 = S_d^2 - S_e^2 = .0082$	$S^2 = S_d^2 - S_e^2 = .0344$
$S = .0906$	$S = .1855$
$S_e^2/S_d^2 = .613$ or 61%	$S_e^2/S_d^2 = .1157$ or 12%
90% credibility interval for d: (-.05, .25)	90% credibility interval for d: (.05, .57)

(Notation is explained in Table 2.)

TABLE 6

META-ANALYSES OF GROUPED TESTS:
POST-TEST VS. PRE-POST DESIGNS WITHIN TEST TYPE

Type 1: High-stakes, admissions tests

Studies with post-test only	Studies with pre- and post-tests
No. of studies = 3	No. of studies = 13
Total N = 1106	Total N = 5674
$\bar{d}$ = .10	$\bar{d}$ = .20
S_d^2 = .0020	S_d^2 = .0181
S_e^2 = .0112	S_e^2 = .0005
$S^2 = S_d^2 - S_e^2$ = -.0092	$S^2 = S_d^2 - S_e^2$ = .0176
S = (.00)	S = .1327
S_e^2/S_d^2 = >100%	S_e^2/S_d^2 = .0276 or 3%
90% credibility interval for d: (.10, .10)	90% credibility interval for d: (-.02, .42)

Type 2: General intelligence tests

Studies with post test only	Studies with pre- and post-tests
No. of studies = 4	No. of studies = 17
Total N = 1049	Total N = 1537
$\bar{d}$ = .10	$\bar{d}$ = .48
S_d^2 = .0414	S_d^2 = .0431
S_e^2 = .0163	S_e^2 = .0160
$S^2 = S_d^2 - S_e^2$ = .0251	$S^2 = S_d^2 - S_e^2$ = .0271
S = .1584	S = .1646
S_e^2/S_d^2 = .394 or 39%	S_e^2/S_d^2 = .371 or 37%
90% credibility interval for d: (-.16, .36)	90% credibility interval for d: (.21, .75)

(Notation is explained in Table 2.)

TABLE 7

EFFECT SIZES OBTAINED BY THREE META-ANALYSES

	Kulik et al.	Pearlman	Witt
All studies	.33	.20	.22 (.23)[a]
Test type			
SAT/High-stakes	.15	.19	.18
Non-SAT/Intelligence	.43	.24	.33
Study design			
Post-test only	.27	.07	.10
Pre- and post-tests	.40	.24	.26
Design within test type			
High-stakes			
Post-test only	---	---	.10
Pre- and post-	---	---	.20
General intelligence			
Post-test only	---	---	.10
Pre- and post-	---	---	.48

Note. Figures in the first and second columns can be compared directly, but figures in the third column were obtained using slightly different sample sizes and test-type classifications. Any comparisons with the third column should be made cautiously.

[a]Effect size after correction for measurement error