Review of Educational Research
Winter 1995, Vol. 65, No. 4, pp. 483–508

Race Comparisons on Need for Achievement: A Meta-Analytic Alternative to Graham's Narrative Review

Harris Cooper and **Nancy Dorr**
University of Missouri–Columbia

A box score review conducted by Graham (1994) concluded that no difference existed between Blacks and Whites on measures of need for achievement. A meta-analysis reported in this article using the same research base revealed reliable and complex race differences. Overall, Whites scored higher than Blacks on measures of need for achievement, but the race difference all but disappeared in studies conducted after 1970. As a possible explanation, the meta-analysis revealed that since 1970 samples of participants from various socioeconomic levels have been preferred and that such samples showed differences between races of only half the size of those shown for samples of participants of strictly lower socioeconomic status. The method of assessment and the age and education of participants also influenced outcomes of race comparisons. Finally, Graham concluded that the research showed a consistent pattern of more positive self-concept of ability among Blacks than Whites. The meta-analysis also found this effect but revealed it to be smaller (though nonsignificantly so) than the difference in need for achievement rejected by the box score. Thus, the meta-analysis found that effects are no larger in an area where Graham concluded they existed than in an area where she concluded they did not.

> Questions of motivation are at the heart of contemporary concerns about the status of African Americans in general and their academic achievements in particular. (Graham, 1994, p. 55)

So begins Sandra Graham's (1994) broad review of research on the academic motivation of African Americans. Graham argued that poor school performance by Black students is often attributed to low personal expectations, feelings of hopelessness, denial of the importance of individual effort, or lack of persistence. Graham went on to contend that whereas poor motivation is frequently invoked to explain disappointing achievement among Black children, studies of Black psychology have all but vanished from mainstream research journals. Research that does exist, she maintained, is largely atheoretical in approach and simplistic in problem formulation, dealing mainly with comparisons between Blacks and Whites.

Support for this project was provided by the Center for Research in Social Behavior, University of Missouri–Columbia. We thank B. Ann Bettencourt and Helen Neville for comments on a draft of this article.

Further, implicit in the prevailing research approach are the untested premises that poor motivation is internal to African Americans (that is, it generalizes over situations and time) and is transmitted intergenerationally through a childhood socialization process.

Graham held up research on need for achievement as the best example of this point of view. She argued that studies of race differences in need for achievement are simplistic and laden with implicit ideology. Most interestingly, her review of research led her to conclude that the presupposition of racial differences was itself erroneous. Graham summarized her findings as follows:

> [A]re data consistent with the belief . . . that African Americans lack the motive to strive for success? As a question that primarily was posed in a comparative racial framework, the studies after 1970 suggest a definitive answer of "No," both when the motive scores were investigated as main effects and when their relations to other variables were examined. (p. 68)

A Role for Research Synthesis

We share many of Graham's concerns. We are especially impressed by the breadth of her review, her attempt to integrate ten lines of research, and her thoughtful commentary on how future research should proceed. Hopefully, her discussion of race and motivation will renew psychologists' interest in the role race plays in American education. However, a renaissance of interest will be short-lived unless new theorizing and research rests on a sound empirical base, the best that can be offered. Reviews of empirical literature, like Graham's, provide one importance source of scientific underpinning.

In the past two decades, procedures for conducting syntheses of empirical research have undergone a dramatic transformation (cf. Cooper & Hedges, 1994). Subjective and haphazard procedures have been replaced by systematic ones that emphasize explicit decision rules, systematic retrieval of information from constituent studies, and quantitative methods for integration of results. Today, methods of review and decision criteria must stand up to thorough scrutiny in light of our new understanding of how sound research syntheses should be conducted.

In the pages that follow, we examine the methods and rules Graham used to draw her conclusions. First, the new standards for research synthesis will be applied to the methods used by Graham, it will be asked whether the inferences she drew might have differed had she applied different statistical criteria. Next, the consistency of her application of methods across topic areas will be examined by a comparison of the research base underlying her conclusion about need for achievement with the research base underlying her conclusion about self-concepts of ability (she reported a general pattern of more positive self-concept among Blacks than Whites). In order to accomplish these ends, we will apply to Graham's research base the procedures of meta-analysis, using quantitative methods to combine and contrast the estimates of relationship strength from independent studies.

First, we need to state our case for why meta-analysis may be the optimal strategy for synthesizing research bases like the one on race differences in need for achievement. Literature reviews can have many different focuses, goals, perspec-

tives, and coverage strategies (cf. Cooper, 1988). For meta-analysis to be appropriate, the synthesis must (a) focus on empirical studies, (b) have the goal of integrating the results of studies so as to create generalizations and set the limiting conditions of the generalizations, (c) employ a neutral perspective (that is, the research is not being mustered to support a particular point of view determined prior to the review), and (d) cover a near exhaustive selection of relevant studies.

Research on race differences in need for achievement is perfectly consistent with this taxonomic description of meta-analysis. In fact, Graham considered the meta-analysis option but chose a more traditional approach. She explained,

> This decision was guided not only by perceived unevenness in the quality of the studies but also by the fact that many of the older investigations were not well-reported and did not contain sufficient information to calculate relevant effect sizes. (p. 60)

These concerns do not necessarily rule out meta-analysis. First, variations in the quality of constituent studies is no more a barrier to meta-analysis than to narrative review. Although, some quantitative synthesists would opt to discard studies of poor quality and work only with a high-quality research base (cf. Slavin, 1986), others (ourselves included) include nearly all conceptually relevant studies and examine quantitatively whether methodological features are related to study outcomes. If study results are related to methodological quality, then conclusions based on good studies are the ones to be believed. In the analyses that follow, we use the same database of empirical studies used by Graham, assuming she would have excluded studies she deemed had no credence.

Second, we took as a challenge to resourcefulness the assertion that older investigations failed to contain sufficient information to calculate effect sizes. While it is true that recent practices have improved the reporting of data somewhat, there are numerous ways to obtain accurate estimates of relations even when the optimal types of information are missing from the report (cf. Rosenthal, 1984).

Components of a Research Synthesis

The three critical components of research synthesis methodology are the literature search, the retrieval of information from research reports, and the procedures for combining results of studies. Our main focus will be on the last of these, but the first two deserve some mention.

The Literature Search

Graham stated that both manual and computerized searches of psychological, educational, and sociological databases were conducted to find studies for the review (see Graham, 1994, p. 57). These are generally considered to be the most useful methods of research retrieval (Cooper, 1987). Graham made no mention of searching journals that were known to publish relevant research or the examination of reference lists in relevant articles. Often, however, research synthesists carry out these procedures but fail to report them.

Graham's search uncovered 133 studies across five topic areas, each using two research paradigms. This is a substantial research base. Nevertheless, unpublished

research was not included in her review. She confined her research base to journal articles and empirical investigations published in books. Thus, according to Graham, a "vast number" (p. 57) of unpublished dissertations and technical reports were excluded.

There are two reasons frequently given for excluding unpublished research. The first is simply that the research base has the potential to become too large and unwieldy for a narrative review. Quantitative syntheses solve this problem by enlisting the aid of the computer to help with the integration of study results. The second is that unpublished research is generally of lesser quality than published research. Not so, in every instance. Researchers often do not submit their studies for publication because publication is not part of their objective. Moreover, research is often turned down for publication for reasons other than quality. Conversely, most researchers would agree that low-quality research is often published. Suffice it to say that there is much overlap when published and unpublished research is distributed along the quality continuum.

With particular regard to research on race and need for achievement, the distinction in quality between published and unpublished research is likely to be relatively narrow. This is because the race variable is easy to validly operationalize, and most studies (even unpublished ones) employ some standardized measure of the achievement need. Quality is more likely to vary in how tests are administered and scored and how data are analyzed. In sum, then, publication status is a bad proxy for quality, and this may be especially true for research areas that compare well-defined groups on standardized instruments.

However, the distinction between published and unpublished research cannot be ignored. It is not so easy to dismiss the fact that published studies tend to be biased toward significant results. Regardless of quality, researchers are less likely to submit nonsignificant results, and editors are less likely to publish them (Coursol & Wagner, 1986; Greenwald, 1975). Publication bias has been demonstrated by modeling its effects analytically (Lane & Dunlap, 1978), by comparing published with unpublished results (e.g., White, 1982), and by testing for the existence of a relation between effect sizes and sample sizes (Light & Pillemer, 1984). All methods indicate that unpublished studies tend to yield smaller effects than published ones.

What effect should Graham's decision to exclude unpublished studies have on her results? In the case of need for achievement, in which she found no relation to race in published literature, a search of unpublished research should add support to her conclusion. In the case of academic self-esteem, in which she concluded that esteem was higher among Blacks than Whites, the addition of unpublished studies might have mitigated her findings.

In the analyses that follow, we have used Graham's database and have not attempted to supplement it. By yoking our research base to hers, we can determine how differences in analytic procedures alone influence conclusions. However, we recommend that future synthesists interested in psychological research on race conduct a thorough search of the unpublished literature. The larger research base thus created should lead to more precise point estimates and should encompass more variations in samples, instruments, procedures, research designs, and conceptualizations. These can be tested as factors that influence the magnitude of relations. This would add to the richness and rigor of reviews.

The Retrieval of Information From Research Reports

We could find no description in Graham's review of an attempt to estimate the reliability with which decisions about the relevance of studies were made or with which information was extracted from individual studies. In this instance, however, we do not see this as a serious problem because the information used in Graham's review involved little inference on the part of the coder (e.g., sample composition, type of measure of socioeconomic status [SES], method of assessment, direction of findings). Stock, Okun, Haring, Miller, and Kinney (1982) found that simple information can be extracted with quite high reliability by a single trained individual. For more complex judgements—say, those involving quality assessments or values calculated from other values—estimates of reliability are more critical (cf. Orwin, 1994).

Of course, our reexamination of Graham's review serves as a reliability check in itself. Again to preserve the equivalence of the research bases, we have tried to be faithful to Graham's judgments and have given her interpretations preference to our own whenever possible. Where we could not find a legitimate way to replicate Graham's judgments about studies, we will so inform the reader.

The Synthesis of Results Across Studies

Box scores or vote counts. Graham employed a *box score* strategy, alternatively called a *vote count*, to determine whether a series of studies supported a particular hypothesis. Specifically, she wrote,

> Using this global box-score method as a way to interpret the hypothesis of racial differences in motive strength, it would be important to examine the data for a consistent pattern of findings over time as well as for evidence that the majority of studies (i.e. at least 50%) find clear differences in the predicted direction. (p. 61)

The vote counting strategy has much intuitive appeal. However, today that strategy has few defenders. One critical problem is that chance alone should produce only about 5% of all reports falsely indicating that a relation exists (we assume here that by "clear differences" Graham means ones that are statistically significant). Therefore, depending on the number of studies, even if the percentage of findings that are positive and statistically significant is well below 50%—or even 34%, if one distinguishes two directions of significant studies as well as null studies—this may still indicate that a substantial relation is present in the population.

Even more troubling, Hedges and Olkin (1980) demonstrated that vote counts have power characteristics that are inversely related to the number of studies contained in the review. When a real effect exists in the population, the more studies a review covers, the less likely it is that a vote count set at 50% or 34% will reject the null hypothesis. Thus, the vote counting strategy could, and often does, lead to the acceptance of the null hypothesis when in fact such a conclusion is unwarranted.

In addition, the vote counting strategy does not differentially weight studies

based on sample size. This is a problem because a study with 100 participants and a study with 1,000 participants are given equal weight. Further, the revealed impact of the treatment in each study is not considered; a study showing a small negative relation and a study showing a large positive relation are given equal weight. For these reasons, the vote count has lost much of its credibility as a means for drawing inferences in research syntheses.

Effect size estimation. Instead of using the vote count method, synthesists interested in drawing inferences from a series of studies now calculate measures of effect, or relationship strength, and use confidence intervals around averaged effect sizes to test the null hypothesis. If an effect size of zero is not contained in the confidence interval, then the null hypothesis can be rejected.

In the case of race differences in need for achievement, the most appropriate effect size metric would be the standardized mean difference, or *d*-index (Cohen, 1988). The *d*-index expresses the difference between two group means in standard deviation (*SD*) units. So, if $d = .20$ this indicates that one group has a mean score two tenths of an *SD* higher than the other group.

The *d*-index is calculated by dividing the difference between the group means by an *SD* estimate. The *SD* estimate can come from one group or the other, or it can be an average of the two groups' *SD*s. The former strategy is typically used when one group serves as a treatment and one as a control, when the *SD*s are suspected to be unequal, and/or when the researcher wishes to express how much of a difference it makes to receive the treatment (the control *SD* is then used in the denominator). We did not use this strategy because neither Whites nor Blacks could be considered controls in this review. Instead, we assumed that the *SD* for Whites and Blacks would be roughly equal. Therefore, the best estimate of either would be the average of the two. The assumption of roughly equal *SD*s also underlies the *t*-test and *F*-test commonly used in each study to test the null hypothesis that White and Black need-for-achievement test scores are equal. The equal *SD*s assumption also permits the estimation of *d*-indexes from *t*s, *F*s, and sample sizes plus *p*-values.

The weighted average effect size across a series of studies is then calculated by multiplying each *d*-index by the inverse of its variance and dividing the sum of these calculations by the sum of the weights (see Hedges & Olkin, 1985, p. 110). Functionally, this procedure gives proportionally greater weight to effect sizes based on larger samples. Weighted average effect sizes are more precise estimates of population values than unweighted ones. A confidence interval is then calculated for this weighted estimate (see Hedges & Olkin, 1985, p. 113).

Influences on effect size magnitude. Graham stated that results must not only pass the box score test but must also show consistency over time to conclude that a racial difference was "unambiguous" (see p. 66, in addition to the passage quoted earlier). Requiring consistency over time is equivalent to requiring that observed relations not be systematically associated with the years in which the studies were conducted. However, Graham did not state how much of an influence time would need to have in order for a result to be considered inconsistent. Indeed, as is typical of narrative reviews, the criteria for the significance of moderating variables (e.g., SES, age) were not explicated, although conclusions were drawn about several such variables.

In meta-analysis, two procedures are used to test whether research results differ

based on characteristics of studies. First, the synthesist can look for studies that contain tests of the moderating variables within their research designs—for example, studies that examine whether race differences in need for achievement are the same for low- and middle-class populations. This evidence, called *within-study evidence* (Cooper, 1989), is especially valuable because it will typically control for some possible confounding variables. For example, within a particular study, test administrators and scorers are likely to be identical or similar for each participant in the study.

The second kind of evidence involves comparisons between studies that vary on a characteristic. For example, the average relation from studies using only lower-class participants can be compared to the average relation uncovered in studies using only middle-class participants. This *between-study evidence* is more equivocal than within-study evidence because other, possibly confounded, study characteristics may not be held constant. For example, researchers who study lower-class populations might choose different measures of need for achievement than do researchers who study middle-class populations. In such a case, we are unable to tell whether the SES difference would still have emerged if the same instrument had been used with both SES groups. (This added equivocality, by the way, is not introduced by meta-analysis but is inherent in the research base, whether combined using effect sizes, using box scores, or narratively.)

Although between-study evidence is more equivocal than within-study evidence, it is nevertheless a major contribution of research synthesis. Often, characteristics of studies vary in important ways that have never appeared—indeed, could never appear—as variables within a study. The between-studies analysis allows a first approximation of what the influence of these variables might be. For example, in the present synthesis a between-studies analysis will allow us to ask whether the outcomes of race comparisons on need for achievement have systematically changed between 1953 and 1983.

In meta-analysis, between-studies evidence is tested using a formal statistical procedure. Because effect sizes are imprecise, they will vary somewhat even if they all estimate the same underlying population value. The procedure, called *homogeneity analysis*, allows the synthesist to test whether sampling error alone accounts for this variation or whether features of studies, samples, treatment designs, or outcome measures also contribute to variation. Much like a primary researcher, the synthesist groups studies according to potentially important characteristics and tests for between-group differences.

The techniques for calculating homogeneity analyses differ depending on whether the relationship under study is conceptualized as a fixed or random variable (see Cooper & Hedges, 1994, and Hedges & Olkin, 1985, for discussion). We will suggest that race difference in need for achievement is a fixed variable, that is, all studies are drawing samples from a unitary population (this assumption appears to underlie Graham's descriptions of her results, as well).

With regard to calculation, homogeneity analysis results in a chi-square statistic that if significant indicates that more variance exists in effect size estimates across studies than predicted by sampling error alone. Homogeneity analysis can be applied to groups of individual effect sizes and to effect sizes averaged across groups of individual estimates. When it is applied to group average effects, significant results can be interpreted as suggesting that the grouping variable is

associated with a statistically significant amount of variance in the individual effects, similar to a significant effect in analysis of variance.

A Meta-Analysis of Race Comparative Studies

Graham reviewed 19 studies that compared Blacks and Whites on need-for-achievement test scores. She described her box score analysis as follows (only references have been omitted from the passage):

> . . . 7 of 19 studies, or 36%, reported Whites to be higher in Nach [need for achievement] than Blacks, whereas 6 investigations revealed no differences in motive strength between the two racial groups. An additional three studies . . . showed higher Nach among Whites for some subset of the population. In contrast, one study . . . unambiguously showed Blacks to have the greater motive strength, and two others revealed partial findings in favor of Blacks . . . . (p. 61)

Graham then summarized her findings:

> Note also that the Veroff and Peele (1969) investigation was the last to report unequivocal evidence that Whites have the greater motive strength. Thus neither of the suggested criteria (i.e., a majority of studies reporting unambiguous racial differences and consistency over time) supports the assumption that African Americans have less achievement motive than do Whites. . . . (p. 66)

What happens when the same research base is subjected to meta-analysis? Table 1 contains supplemental descriptions of each of the 19 research reports included in Graham's Table 2 (p. 62). There are a total of 26 entries in our Table 1 because in some instances a single research report contained more than one study or a single study reported separate results for independent subsamples of participants. Also, a total of 36 *d*-indexes were calculated because the same participants were given more than one measure of need for achievement, or because the measure of need for achievement had multiple subscales, or because the same participants completed the same measure on more than one occasion. When more than one *d*-index was calculated for a single independent sample, these were averaged within samples so that effect sizes remained independent (see Cooper, 1989, pp. 76–79, for an explanation of this procedure). Because average Black need-for-achievement scores were subtracted from average White need-for-achievement scores, positive values in Table 1 indicate that Whites average higher scores than Blacks.

Note first that we were able to obtain an effect size for every relevant race comparison in every study. Table 1 lists the method used to derive each effect size. Most frequently, effect sizes were calculated using reported means and standard deviations or by converting *t*-tests or *F*-tests to *d*-indexes (the *d*-index equals 2 times the *t*-value divided by the square root of the degrees of freedom for error; cf. Rosenthal, 1984). In a few instances, we derived the *t*-value from reported sample sizes and *p*-levels. In three instances, we generated our own statistics based on frequencies given in the tables of the primary reports.

Figure 1 presents a stem and leaf display of the 26 *d*-indexes for independent

TABLE 1
Studies of race comparisons on need for achievement

Source of independent sample	Effect size	Number of effect sizes	Total *N*	Nach measure	Age/education of subjects	SES	Data used to calculate effect size
Mussen (1953)	+.85	1	100	TAT pictures	Elementary & junior high	Low	Test value
Rosen (1959)	+.38	1	427	TAT pictures	Elementary & junior high	Varied	Significance level
Veroff et al. (1960), male subjects	+.02	1	569	TAT pictures	Adults	Varied	χ^2 converted to *d*
Veroff et al. (1960), female subjects	+.10	1	748	TAT pictures	Adults	Varied	χ^2 converted to *d*
Smith & Abramson (1962)	.00	1	66	TAT pictures	High school	Low	Significance level
Lott & Lott (1963)	+.55	1	253	Sentence as stimulus	High school	Varied	Test value
Mingione (1965), Study 1	+.34	1	105	Altered TAT drawings	Elementary & junior high	Low	Test value
Mingione (1965), Study 2	+.35	1	245	Altered TAT drawings	Elementary, junior high, & high school	Low	Test value
Mingione (1968), female 5th graders	+.05	1	40	Sentence as stimulus	Elementary	Low	Means & *SD*

TABLE 1 *(continued)*

Source of independent sample	Effect size	Number of effect sizes	Total *N*	Nach measure	Age/education of subjects	SES	Data used to calculate effect size
Mingione (1968), male 5th graders	+.30	1	46	Sentence as stimulus	Elementary	Low	Means & *SD*
Mingione (1968), female 7th graders	−.10	1	72	Sentence as stimulus	Junior high	Low	Means & *SD*
Mingione (1968), male 7th graders	+.01	1	58	Sentence as stimulus	Junior high	Low	Means & *SD*
Baughman & Dahlstrom (1968)	+.44	1	480	Altered TAT drawings	Elementary	Low	Test value
Garza (1969)	+.48	1	129	TAT pictures	Elementary	Low & middle	Generated means & *SD* based on values in table
Veroff & Peele (1969)	+.28	2	1,390	Behavioral	Elementary	Not specified	Test value
Turner (1972)	+.30	1	518	TAT pictures	Junior high	Low & middle	Significance level
McClelland (1974)	−.02	2	365	TAT & behavioral	Adults	Not specified	Test values
Hall (1975)	.00	1	159	Sentence as stimulus	College	Low & middle	Significance level

Travis & Anthony (1975)	–.72	1	135	Forced choice	High school	Low & middle	Test value
Ramirez & Price-Williams (1976), female subjects	+.11	2	60	Altered TAT drawings	Elementary	Low & middle	Means & *SD*
Ramirez & Price-Williams (1976), male subjects	–.02	2	60	Altered TAT drawings	Elementary	Low & middle	Means & *SD*
DeBord (1977), underachievers only	+.40	1	48	TAT pictures	Elementary	Low	Means & *SD*
DeBord (1977), achievers only	+.17	1	45	TAT pictures	Elementary	Low	Means & *SD*
Ruhland & Feld (1977)	+.04	2	193	Sentence as stimulus	Elementary	Low	Test value
Leftkowitz & Fraser (1980)	–.02	3	63	TAT pictures & Guttman scaled measure	College	Varied	Means & *SD*
Castenell (1983)	–.33	4	310	Forced choice	Junior high	Low & middle	Test value

Note. Positive effect sizes indicate higher mean for White sample than for Black sample. Nach = need for achievement; SES = socioeconomic status; TAT = Thematic Apperception Test.

TABLE 2
Mean d*-indexes for White-Black comparisons on need for achievement*

Moderating variable	*k*	*d*-index	Confidence interval
Overall[a]	26	+.21	+.15, +.27
Year[b]			
Before 1970	15	+.32	+.25, +.39
After 1970	11	+.02	−.07, +.11
SES[c]			
Lower	12	+.31	+.21, +.41
Lower, middle, and varied	12	+.15	+.06, +.24
Type of Nach measure[d]			
TAT pictures	9	+.29	+.19, +.39
Altered TAT pictures	5	+.36	+.23, +.49
Nonpictorial stimuli	10	+.11	+.02, +.20
Age/education[e]			
Elementary school	10	+.29	+.20, +.38
Junior high school	4	+.10	−.04, +.24
Senior high school	3	+.09	−.10, +.28
College	2	−.004	−.27, +.26
Adult	3	+.03	−.12, +.18
Elementary school only[f]			
Lower SES	6	+.31	+.17, +.45
Lower, middle, and varied SES	3	+.27	+.02, +.52
Junior and senior high only[g]			
Lower SES	3	−.03	−.32, +.26
Lower, middle, and varied SES	4	+.13	+.01, +.25

Note. Positive values indicate higher mean for White sample than for Black sample; k = number of samples from separate studies contributing to estimate.
[a]$\chi^2(25) = 82.32$, $p < .001$; [b]$\chi^2(1) = 24.67$, $p < .001$; [c]$\chi^2(1) = 5.47$, $p < .02$; [d]$\chi^2(2) = 11.12$, $p < .005$; [e]$\chi^2(1) = 10.98$, $p < .001$; [f]$\chi^2(1) = 0.07$, n.s.; [g]$\chi^2(1) = 0.97$, n.s.

samples. As reported in Table 2, the unweighted average *d*-index equaled +.15 (median $d = +.10$). The weighted average *d*-index equaled +.21 with a 95% confidence interval ranging from $d = +.15$ to $d = +.27$. Thus, while the 50% box score method suggested the overall evidence was ambiguous, the effect size analysis revealed that Whites had higher need-for-achievement scores than Blacks. Is a difference of two tenths of a standard deviation small or large? We will return to this question after describing additional outcomes of the meta-analysis.

Consistency Over Time

What about consistency over time? First, a homogeneity analysis revealed that there was considerably more variability in the 26 individual *d*-indexes than would be predicated by sampling error alone, $\chi^2(25) = 82.32$, $p < .001$. To examine this variation further, we grouped studies into those published prior to 1970 and those published after 1970 (the year 1970 was chosen because Graham used this date to distinguish early from recent studies). The results of this between-studies analysis

.8	5
.7	
.6	
.5	5
.4	0 4 8
.3	0 0 4 5 8
.2	8
.1	0 7
.0	0 0 1 2 2 4 5
−.0	2 2
−.1	0 5
−.2	
−.3	3
−.4	
−.5	
−.6	
−.7	2

FIGURE 1. *Stem and leaf display of* d*-indexes for race comparisons on need for achievement*

revealed that the year in which a study was published predicted a significant amount of variance in effects, $\chi^2(1, N = 26) = 27.8$, $p < .001$. As the effect size averages in Table 2 show, the difference is dramatic. In the earlier studies, need-for-achievement scores for Whites were three tenths of a standard deviation higher than for Blacks. More recent studies revealed no race difference whatsoever. Thus, the meta-analysis confirms Graham's conclusion.

But what explains the diminishing difference? Graham used the temporal effect to infer that no relation existed between need for achievement and race. We suggest the effect of time itself requires further explanation. That is, time likely serves as a proxy for a more substantive underlying process description. For instance, perhaps changes in culture have acted to minimize racial differences by lowering White need for achievement, raising Black need for achievement, or both. Regrettably, the present data set provides no way to test such a hypothesis.

We should ask also whether the methods of studying race differences in need for achievement have changed over time. That is, can co-occurring changes in instruments, research procedures, or sampling designs account for the temporal trend? Our approach was to search for co-occurring changes by correlating the year in which a study was published with three other characteristics of studies: (a) the SES of participants, (b) whether the study employed Thematic Apperception Test (TAT) pictures as the method of assessment (versus TAT pictures altered to reduce possible racial bias, versus measures using nonvisual stimuli), and (c) the age (or educational level) of the participants. The first two variables were also examined by Graham, but not in the context of explaining the temporal trend in the data. We then applied the within- and between-study analyses of moderator variables described above.

Socioeconomic status. The correlation between the year in which a study was published and the socioeconomic makeup of the sample it contained ($r(19) = .51$, $p < .03$) indicated that more recent studies were more likely to sample higher-SES

participants. For example, prior to 1970 nine samples were described by researchers as composed solely of lower-class individuals, and five were described as composed of lower- and middle-class individuals or as containing participants of various SES levels. After 1970, three samples were described as lower class and seven were described as lower and middle class or varied. (In two studies the SES of the participants was not specified).

Given the substantial relation between the SES of participants and year of publication, it is important to ask two questions. First, the main effect question: Does SES relate to need for achievement? Second, the interaction or interdependence question: Is the relation between race and need for achievement dependent on the SES of the participants?

Graham found five studies that examined one or both of these questions. For three studies (Castenell, 1983; Rosen, 1959; Smith & Abramson, 1962), Graham reported that higher motive scores were associated with increasing SES. We were able to retrieve an effect size only from the Castenell (1983) study. We agree about the direction of findings in the Rosen (1959) study, but we could find no test of the SES main effect in Smith and Abramson (1962). Also, we found two additional tests of the main effect in the current research base (Hall, 1975; Leftkowitz & Fraser, 1980).

The average weighted *d*-index across these two studies and Castenell (1983) was $d = +.05$ with a 95% confidence interval ranging from $-.13$ to $+.23$. It should be pointed out that the two unusable studies were conducted before 1970 and all three studies included in the effect size analysis were conducted after 1970.

Regardless, the main effect results are of limited use for understanding the temporal disappearance of the race difference. If the SES of participants in race comparative research has increased over time (which it has), a main effect of SES on need for achievement suggests only that the *overall* average need for achievement scores, regardless of race, should be on the rise. The main effect does not address directly the magnitude of the race comparison. To address this question, it is necessary to ask whether the race difference is consistent across all SES levels. If it is not—specifically, if it is larger among lower-class than middle-class samples—then this fact in tandem with the upward shift in sampled SES levels over time could explain the diminishing relationship.

Graham cited three studies that contained evidence about the moderating effect of SES on differences between races in need for achievement. In one study, Rosen (1959) found a significant main effect for both race (comparing Blacks to five White ethnic groups) and social class. A significant interaction was also found. Inspection of the means underlying the interaction revealed that Blacks had higher need-for-achievement scores in the highest SES group whereas Whites had higher need-for-achievement scores in the three remaining SES groups. In another study, Lott and Lott (1963) found no differences in need for achievement among a small subsample ($n = 30$) of participants matched on both SES (all were middle class) and IQ scores (all between 85 and 104). Finally, Turner (1972) reported no differences between Blacks and Whites in a group of students whose fathers were farm subsistence laborers (White $n = 13$, Black $n = 89$) but significantly higher need for achievement for Whites among students whose fathers were manual unskilled laborers (White $n = 48$, Black $n = 77$) or manual skilled laborers (White $n = 182$, Black $n = 29$).

In characterizing the literature as a whole, Graham said the studies are few and varied in method (we agree) but also that "the general pattern of findings is consistent" and that "SES, when examined, has been at least as important a factor as race in accounting for group differences in need for achievement" (p. 67). Thus, Graham did not make a clear distinction between studies that tested SES as a main effect and those that tested SES as a moderator of race differences. If her conclusion related to tests of SES as a moderator variable (as her use of the phrase "accounting for group differences" might imply), we initially found the within-study evidence inconsistent. For middle-class participants, one study revealed higher mean scores for Blacks, one revealed higher mean scores for Whites, and one revealed no difference. For lower-class participants, one study revealed higher mean scores for Whites and one study revealed one finding of a higher mean for each group.

A meta-analysis of between-study differences can provide a preliminary test of the race and SES interaction. We divided the 26 independent samples into those that were described by the researchers as containing only lower-class participants (k = 12) and those that contained both lower- and middle-class participants or varied SES levels (k = 12).

As Table 2 displays, the race differences among lower-class samples averaged about three tenths of a standard deviation whereas the race difference among the mixed-class samples was significantly lower and about half the size, $\chi^2(1, N = 24) = 5.47$, $p < .025$.

Reexamining the within-study evidence in light of these findings revealed that only one small subsample (Turner's [1972] subsistence farmers, based on a comparison involving a sample of 13 Whites) truly contradicted the between-study analysis. In sum, then, the meta-analysis suggested that race differences in need for achievement might be larger in lower-class than in mixed-class samples. Therefore, the more recent trend toward using higher-SES samples is one possible explanation for the reduction in race differences in need for achievement over time.

Method of assessment. Unlike SES, the method of assessing need for achievement (specifically, whether or not TAT pictures were used) was found to be unrelated to the year in which a study was conducted, $r(24) = .37$, n.s. However, examining the method of assessment as a potential moderator of race differences remains of interest to researchers. Specifically, we need to consider whether attempts to remove possible race-related bias from need-for-achievement instruments influence the outcome of race comparisons.

Graham located six studies conducted after 1950 that compared the TAT responses of Black participants presented with White and Black stimulus characters in TAT pictures but found that these studies focused on the length of written protocols instead of on the number of achievement themes. The only study that examined need-for-achievement scores (Cowan & Goldberg, 1967) found a small and nonsignificant effect (d = +.03) indicating higher scores for Blacks when stimulus figures were Black.

In addition, we found two studies that compared assessment procedures involving two types of stimuli: pictorial and nonpictorial (e.g., sentence stimuli, behavioral, Guttman-scale). McClelland (1974) found nearly identical outcomes on the two types of measures. Leftkowitz and Fraser (1980) found that TAT pictures

(both adjusted and unadjusted) led to higher scores for Whites than for Blacks (d = +.20) whereas a 29-item Guttman-scale self-report measure found higher scores for Blacks than for Whites (d = –.43).

To conduct a between-studies analysis, we divided the samples into three groups that used (a) traditional TAT pictures (k = 9), (b) TAT pictures altered to disguise the race of the stimulus characters (k = 5), and (c) nonpictorial stimuli (k = 10). The analysis revealed significant differences between the average effect sizes for the three types of instruments, $\chi^2(2, N = 24) = 11.12$, $p < .005$. As Table 2 displays, samples using TAT pictures revealed an average race difference of d = +.29, those using altered TAT pictures had an average d = +.36, and those using nonpictorial measures had an average d = +.11. Single degree of freedom contrasts indicated that using unaltered and altered TAT pictures did not influence the magnitude of the race difference, $\chi^2(1, N = 14) = 0.53$, n.s., but using nonpictorial scales yielded smaller race differences than using either altered or unaltered TAT pictures, $\chi^2(1, N = 24) = 10.59$, $p < .005$.

These results support Graham's statement that "there is no clear evidence . . . that the motive scores of African Americans are significantly influenced by the racial characteristics of stimulus persons" (p. 61). However, both within- and between-study evidence indicated that TAT measures obtained larger race differences than did measures not employing pictorial stimulus materials.

Age or education of participants. The final moderator examined in the meta-analysis concerned the age or education level of the participants (age and education level were nearly perfectly confounded variables). The age/education level of the sample was not significantly related to the age of the study, $r(22) = -.30$, n.s. Therefore, as for assessment methods, change in sample age is an unlikely explanation for the change in race comparison results over time.

There were four individual studies that included samples drawn from distinct age groups. Two of these (Mussen, 1953; Rosen, 1959) did not include age in any analyses, nor did they present age-related means. Mingione (1965) found no significant age-race interaction, but apparently employed age groupings as a categorical rather than continuous variable. Mingione (1968) also found a nonsignificant age-race interaction; the means indicated that Whites scored higher in fifth grade and Blacks in seventh grade.

Between studies, we found 10 independent samples of elementary school students (kindergarten through Grade 6), 4 samples of junior high school students (Grades 7 and 8), 3 of high school students (Grades 9 through 12), 2 of college students, and 3 of noncollege adult populations (4 samples cut across our age groupings) in which Whites and Blacks were compared on need for achievement.

A homogeneity analysis revealed significant variation in the average d-indexes related to age/education level, $\chi^2(1, N = 22) = 10.98$, $p < .001$, indicating that larger differences between Whites and Blacks were found in younger samples. As Table 2 shows, for elementary school students, the race difference was about three tenths of a standard deviation; for junior and senior high students the difference was about one tenth of an *SD*; no difference was found for college and noncollege adults.

Because differences were related to the age of the sample, it was important to again consider whether this variable was confounded with the other moderator variables. First, whereas there was some confounding between the method of

assessment and the age of the sample, it was not linear in effect; 70% (7 of 10) of elementary school samples were given TAT-type measures, 29% (2 of 7) of junior and senior high school samples were given TAT-type measures, and 60% (3 of 5) of adult and college samples were given TAT-type measures. For SES, the confounding was decidedly more linear; 67% (6 of 9) of elementary school samples were lower class, 43% (3 of 7)of junior and senior high school samples were lower class, and 0% (0 of 4) of college and adult samples were lower class.

The confounding of age and SES suggested that the race differences in elementary school participants should be tested for constancy across samples of different SES. The analysis showed the difference to be nonsignificant, $\chi^2(1, N = 9) = .07$, n.s. For lower-class elementary school samples the average *d*-index was +.31 (with a 95% confidence interval from +.17 to +.45) whereas for mixed-class elementary school samples the average *d*-index was +.27 (with a 95% confidence interval from +.02 to +.52). No difference due to SES was found for junior and senior high school samples.

A Comparison With Self-Concept of Ability

As a final exercise in distinguishing the box score and meta-analytic methods, we examined the research base from which Graham drew her conclusion that Blacks had higher self-concepts of ability than Whites. In her interpretation of the research, she wrote,

> Of the 18 studies listed, only the 2 earliest investigations . . . unambiguously report that Whites had higher academic self-esteem than did Blacks. Thereafter, 7 studies showed Blacks to be higher in self-perceived ability, 4 revealed no significant differences between the races, and 5 reported mixed results (i.e., at least partial findings in favor of African Americans). (p. 98)

On this basis, Graham concluded there was a "general pattern of higher self-concept among African Americans" (p. 98). Does an analysis of effect sizes suggest a similar or different conclusion, and how does the magnitude of this relation compare with that revealed in the meta-analysis of race and need for achievement?

For this meta-analysis the scores of Whites were subtracted from those of Blacks; therefore, positive effect sizes mean that Blacks' self-concept was higher than Whites'. Of the 18 studies listed in Graham's Table 7, 6 were excluded from our analysis. In 4 instances, data were insufficient to compute a *d*-index. Four of these studies (Kugle, Clements, & Powell, 1983; Nichols & McKinney, 1977; Wylie, 1963; Wylie & Hutchins, 1967) reported either significantly higher self-concepts for Whites or nonsignificant differences between the races (one reported higher scores among Whites, two among Blacks). Excluding them from the sample should serve only to increase the average *d*-index, if the included studies in fact reveal higher self-concept of ability among Blacks. A study by Drury (1980), which favored Whites overall but favored Blacks when SES was controlled, was not included because it used the school instead of the participant as the unit of analysis. Finally, a study by Hunt and Hunt (1977) was excluded because it was a reanalysis of data collected for another study already in the

sample (Rosenberg & Simmons, 1971).

Our computations were consistent with Graham's interpretation for 8 of the 12 studies that yielded effect sizes. For the other 4 studies included in the meta-analysis, we had to make questionable assumptions in order to arrive at Graham's interpretation of results, all of which favored her conclusion. In 2 studies, overall relations showing higher self-concept among Whites were accompanied by additional analyses indicating that when a third variable was controlled higher self-concept was found among Blacks (in Lay & Wakstein, 1985, achievement was controlled; in Gray-Little & Appelbaum, 1979, IQ was controlled). We included only the *d*-indexes based on the analyses with controlled variables. For Olsen (1972), Graham reported that the ability self-concept scores were more positive for Blacks than Whites. However, Olsen tested for differences within race over time. In order to replicate Graham's review, we made the assumption that the race difference was also significant and computed a *d*-index based on sample size and $p < .05$. Finally, Graham reported that academic ability self-concept scores were greater for Blacks than Whites in Rosenberg and Simmons's (1971) data. Our inspection of their results—produced by asking participants the question, "How smart do you think you are?" (Rosenberg & Simmons, 1971, p. 94)—revealed a difference favoring Whites. Therefore, we assumed that Graham used a general self-esteem measure on which scores for Blacks were more positive than scores for Whites. We included only this effect size in the analysis.

The 12 studies yielded 14 *d*-indexes from independent samples. The average weighted *d*-index of +.16 (with a 95% confidence interval from +.14 to +.18) indicated that Blacks reported more positive self-concepts of ability than Whites. The unweighted average *d*-index equaled +.13 and the median was $d = +.13$. Also, Graham suggested that more recent studies showed larger race differences. The meta-analysis revealed that the average *d*-index reported between 1969 and 1978 was greater in magnitude ($d = +.37$, $k = 7$) than that reported between 1979 and 1985 ($d = +.11$, $k = 7$), $\chi^2(1) = 48.92$, $p < .001$.

Finally, we found no significant difference between the absolute magnitude of relation between race and need for achievement ($d = .21$) and the absolute magnitude of relation between race and self-concept of ability ($d = .16$), $\chi^2(1, N = 40) = 2.69$, n.s. Thus, while Graham claimed no difference between the races on need for achievement and reliably higher scores among Blacks on self-concept of ability, the meta-analysis revealed that (a) even when many disputable points are granted to Graham, the research on self-concept of ability is indistinguishable from that on motivation to achieve, and (b) the self-concept relation, like the need for achievement relation, appears to have diminished with time, although not as dramatically.

Comparing the Meta-Analytic and Narrative Results

In sum, the narrative and box score review of need for achievement concluded that no unambiguous race difference existed, and that this finding was most clear in studies conducted in the past 25 years. The meta-analysis found clear evidence that a race difference existed ($d = +.21$) but agreed that it had diminished to near zero in studies conducted after 1970 ($d = +.02$).

Graham appeared to imply that when SES was held constant race differences

were not found. The meta-analysis found race differences among lower-class samples to be both substantial ($d = +.31$) and twice that of samples containing both middle- and lower-class participants ($d = +.15$). Within-study evidence of race differences in strictly middle-class samples did not permit quantitative analysis but did suggest that differences were quite small, if they existed at all. The meta-analysis also suggested that a clue to the cause of the diminishing race difference over the years might be contained in the simultaneous diminishing of lower-class samples in research.

The narrative review found little evidence that altering TAT pictures made a difference to the results of race comparisons. The meta-analysis supported this conclusion. However, it also revealed evidence that measures of need for achievement that did not rely on visual stimuli revealed smaller race differences ($d = +.11$) than TAT or altered-TAT pictorial tests ($d = +.32$).

Finally, the age/education of the sample may influence race differences in need for achievement. The meta-analysis found this variable to be a significant predictor of variance in race comparisons; younger samples showed significantly larger race differences.

It is no surprise that Graham's box score review found no race differences in need for achievement whereas the present meta-analysis revealed a reliable and complex set of relationships. First, as noted above, box scores hold evidence to a highly conservative standard for acceptance, one that could be expected to lead to frequent Type II errors. Second, research has demonstrated that narrative reviews in general lead to the underestimation of the reliability and magnitude of effects (Cooper & Rosenthal, 1980).

Somewhat more surprising is that the box score analysis led to the conclusion of a reliable race difference in self-concept of ability when the difference appears to be no more trustworthy than that for achievement motivation. Why did this happen? A summary in Table 3 of the box score descriptions of the two research areas suggests that the rule that a "majority of studies (i.e., at least 50%) find clear differences in the predicted direction" (p. 61) was applied to the achievement research (7 of 19 studies) but not to the research on self-concept of ability (7 of 16 studies, including 1 supportive study counted twice).

TABLE 3
A summary of Graham's (1994) box score description of two research areas

Direction of finding	Need for achievement	Ability self-concept
Whites > Blacks	7	2
Partial support of Whites > Blacks	3	0
No difference	6	4
Partial support of Blacks > Whites	2	5
Blacks > Whites	1	7

Interpreting the Effects of Moderating Variables

Readers interested in a set of issues and findings remarkably parallel to race and need for achievement are referred to the literature concerning race difference on the MMPI personality inventory (Dahlstrom, Lachar, & Dahlstrom, 1986; Greene, 1987; Gynther, 1989). In this personality domain, data also appear to present evidence of cultural bias in testing and diminishing race differences with increasing SES and age.

Before concluding our remarks on the value of meta-analysis for the integration of research literatures, we would like to briefly interpret the findings concerning the three moderators of race differences in need for achievement. Finally, we will show some ways the dual questions of "Are the uncovered race differences small or large?" and "Why study race difference at all?" might be addressed.

Socioeconomic status. The meta-analysis results leave plausible the possibility that moderate differences exist between lower-class Whites and Blacks on need for achievement but that the difference diminishes markedly with increasing levels of SES. Proposing motivation as the causal variable, Rosen (1959) proposed an explanation for the lack of differences between middle-class Whites and Blacks—indeed, for why middle-class Blacks might have higher need for achievement. He wrote, "This relatively high score for [higher-SES Blacks] indicates, perhaps, the strong motivation necessary for negroes [*sic*] to achieve middle class status in a hostile environment" (p. 53). Unfortunately, this explanation remains as viable today as it was 35 years ago.

In addition, it should be pointed out that studies with strictly lower-class participants do not carry the same assurance of SES equivalence across racial groups as do studies that explicitly match lower-class White and Black participants on SES variables. That is, while all participants in the two samples might be lower class, one sample could still be more economically disadvantaged than the other. Thus, it is possible that the difference in need for achievement between lower-class Whites and Blacks found here is still, in whole or in part, an SES difference masquerading as a race difference.

Finally, the SES main effect on need for achievement found in the present meta-analysis ($d = +.05$) is considerably smaller than that found in an earlier meta-analysis (unweighted $d = +.33$; Cooper & Tom, 1984). Compared to the present meta-analysis, the earlier meta-analysis covered more research because it included studies that examined SES without examining race simultaneously.

Method of assessment. The current meta-analysis also leaves plausible the contention that at least a portion of the race difference in need for achievement is attributable to cultural bias in the TAT, the predominant instrument of measurement in this data set. However, as Graham correctly asserts, evidence of bias created by the content of TAT pictures remains largely untested. Somewhat more compelling is the finding of diminished race differences when instruments that employ no visual images are used. Thus, it is still possible that assessment methods that are more race neutral produce the smallest race differences.

Age or education level. Before a substantive interpretation of the linear effect of age/education on race comparisons is offered, it is important to ask whether the finding might be an artifact of sampling procedures. For instance, it would appear that the youngest samples used in research on need for achievement are most

inclusive, in that nearly all children attend elementary school. High school differences in need for achievement might be restricted somewhat by differential dropout rates associated with both race and achievement striving. College samples are probably most selective in that college attendance is at least partly related to strong need for achievement, regardless of race. Thus, the diminishing race difference related to these three age/education levels is consistent with the alternative explanation that the age/education of the sample is confounded with increasing restrictions on samples and that participants with higher needs for achievement are therefore overrepresented within each racial group.

How do junior high school and noncollege adult samples fit the sampling restriction explanation? Junior high school samples revealed race differences equal to those shown for high school samples. However, the sampling restriction hypothesis would predict less impact of dropping out on junior high school comparisons. Likewise, the noncollege adult samples (McClelland, 1974; Veroff, Atkinson, Feld, & Gurin, 1960) both appear to have been created with credible sampling procedures. Yet, the result drawn from these samples differs little from that drawn from college students.

In sum, a differential sampling explanation for the age/education difference in comparisons should not be accepted at present because the data do not conform perfectly to it. However, strong substantive interpretation of the age/education findings also should be eschewed until studies involving participants from multiple age/education levels who are sampled using equivalent procedures are conducted.

Gauging the Importance of the Race Difference

Are the uncovered race differences small or large? Cooper (1981) suggested that the substantive interpretation of effect sizes can involve three yardsticks: (a) contrasting effects in related domains of interest, (b) practical significance, and (c) research methodology. We will focus here on the first yardstick.

How does the race difference compare to other effects in education? Lipsey and Wilson (1993) presented a summary of 115 meta-analyses that had been conducted in the field of education. These research reviews differed from the present one in that they all examined attempts to intervene in the schooling process, such as instructional practices, classroom organization, and test taking. Cooper, Dorr, and Bettencourt (1995) found the mean *d*-index across all 115 meta-analyses gathered by Lipsey and Wilson to be $d = +.47$ ($SD = .29$). In all, 35% of the meta-analyses revealed effect sizes smaller than that found in the present synthesis ($d = .21$).

Fraser, Walberg, Welch, and Hattie (1987) undertook a similar synthesis of syntheses. They summarized 134 meta-analyses organized into seven factors that influenced achievement (school, social, instructor, instruction, pupil, methods, and learning strategies). The overall average correlation between the seven factors and achievement was $d = .41$ ($SD = .30$). The 25 meta-analyses that dealt with pupil differences (affective, cognitive, physical, disposition to learn) revealed an average *d*-index of .49 ($SD = .36$). Finally, a summary of 92 meta-analyses relating the factors to affective outcomes of schooling revealed an average *d*-index of .22 (*SD* not given).

Numbers can go no further in helping to answer the question, "Are the uncov-

ered race differences small or large?" In the context of the two syntheses of syntheses, we suspect that considering $d = +.21$ as small but not inconsequential would obtain some agreement. However, drawing an analogy to the widely accepted standard test for optimism, a meta-analysis can tell how many ounces of water there are in a glass, other meta-analyses on related topics can tell how large the glass is, but a judgment of whether the glass is more empty or more full will always interact with the perspective of the observer.

Why study race differences at all? We have saved for last what should perhaps have been the first question. We did so because addressing the question "Why study race differences at all?" necessarily involves the introduction of material that might divert attention from the methodological comparison that is the focus of this article.

Several arguments can be made for the abandonment of race comparative studies. First, it can be argued (a) that studies comparing different people to one another is characteristic of the behavior of Europeans, who use themselves as a standard, (b) that Europeans can no longer be used as the standard against which the psychology of other people is judged, and (c) that comparisons are proper only when racial groups are equated on all relevant variables, especially that of culture. According to Azibo (1988), comparative research can commit a "transubstantive error, taking the cultural and psychological norms of one group and applying them in establishing the meaning of the cultural and psychological functioning of another group" (p. 23). If this line of reasoning is suggested to apply to research on need for achievement, it follows that nothing of value is to be learned from a synthesis of comparative research, regardless of whether the synthesis is carried out using meta-analytic, box score, or narrative methods.

Second, as Graham (p. 103) argues, race comparative studies can perpetuate false negative stereotypes instead of highlighting the variation in behavior that exists within racial groups. No doubt this is not just a possibility but a reality. In the statistical lexicon, a false negative stereotype is perpetuated when differences are overgeneralized, when main effects are highlighted even though interactions are known to exist.

The counterargument would be that social scientists are obligated to separate fact from fabrication, regardless of the topic under scrutiny. Further, the surest way to combat overgeneralization is to secure rigorous data that suggest otherwise. Thus, the most potent response to the false stereotype that Blacks are not as motivated to achieve as are Whites would seem to be to note that, in fact, at least 42% of the Black population has a stronger need to achieve than the average member of the White population (based on $d = +.21$), that possible cultural bias in how achievement motivation is defined (perhaps conceptually, perhaps operationally) may exaggerate perceptions of race differences, and that race differences may cease to exist entirely when Blacks enjoy the same economic benefits as middle-class Whites. This response rests firmly on the best scientific evidence we have.

Third, it can be argued that the documentation of race differences on psychological variables is really no explanation at all. That is, it can be argued that the construct of race is theoretically bankrupt and serves only as a proxy for the true causal mechanisms that create variation in human behavior. To place the argument in causal modeling terms, we are interested in achievement motivation

because it appears to be an important precursor to the outcome variable of ultimate interest, namely, achievement. Is race then a critical precursor of achievement motivation?

We agree with Graham (p. 104) that if an affirmative answer to this question is the end point of race comparative studies then the research does more harm than good. However, race can also be viewed as a intermediate link in the sequential chain leading to achievement. In this conceptualization, race differences have important precursors of their own.

The nature of the search for precursors to race differences can vary according to unit of analysis; the search can focus on biological, psychological, or sociological explanations. Thus, some would argue that race is a critical intermediary because it is linked to biological differences among people, and that Whites and Blacks differ in innate physical or mental capabilities. This contention is presently without substantiation and will remain so until skin color can be randomly assigned at birth and/or until a motivation gene is found that is linked to a skin color gene. Further, the meta-analytic evidence showing that the race difference varies as a function of both the operationalization of motivation and the social class of participants provides empirical evidence against any biological explanations.

Others would argue that race is critical because it is linked to psychological and social psychological patterns within individuals and families. Graham thoughtfully outlined this approach in her detailed discussion of principles of a motivational psychology for African Americans.

Still other social scientists (including many psychologists) would argue that race is important because it is linked to inequities in American society that systematically alter the life chances of African Americans. From this viewpoint, racial inequities that have existed in this country for centuries continue to exist today and have important implications for individual human functioning.

The Contribution of Meta-Analysis

Meta-analysis is not a perfect solution to problems in research review. Most of the strengths and shortcomings of the method have been detailed elsewhere (Cooper, 1990), and so we briefly mention only a few here. First, meta-analysis cannot establish whether causal relations exist between variables examined at the level of study differences. Second, meta-analysis cannot overcome the problem of confounding study-level variables. Third, and most important, as Wachter and Straf (1990) point out, meta-analysis is no substitute for wisdom. A statistical method cannot tell us what topics are important to study. A statistical method cannot generate theories that do not already exist. And even less ambitiously, a statistical method cannot point out to its user what variables should be examined as moderators of relationships. Only the human intellect can do these things.

Still, meta-analysis can make critical contributions to the social science enterprise. First, meta-analysis makes explicit the standards of proof being employed by the synthesist. Second, it requires research synthesists to use the same standards of rigor required of the primary researchers upon whose work their syntheses are based. Third, it insures that standards of proof are identical across topic areas, both within and between reviews. Without these characteristics, research

syntheses and the empirical investigations they come to replace lose their unique claim to legitimacy.

References

Azibo, D. (1988). Understanding the proper and improper usage of the comparative research framework. *Journal of Black Psychology, 15,* 81–91.

Baughman, E. E., & Dahlstrom, W. G. (1968). *Negro and white children: A psychological study in the rural south.* New York: Academic Press.

Castenell, L. A. (1983). Achievement motivation: An investigation of adolescents' achievement patterns. *American Educational Research Journal, 20,* 503–510.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Erlbaum.

Cooper, H. (1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology, 41,* 1013–1018.

Cooper, H. (1987). Literature searching strategies of integrative research reviews: A first survey. *Knowledge, 8,* 372–383.

Cooper, H. (1988). The structure of knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1,* 104–126.

Cooper, H. (1989). *Integrating research: A guide for literature reviews* (2nd ed.). Newbury Park, CA: Sage.

Cooper, H. (1990). Meta-analysis and the integrative research review. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 142–163). Newbury Park, CA: Sage.

Cooper, H., Dorr, N., & Bettencourt, B. A. (1995). Putting to rest some old notions about social science. *American Psychologist, 50,* 111–112.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cooper, H., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87,* 442–449.

Cooper, H., & Tom, D. Y. H. (1984). Socioeconomic status and ethnic group differences in achievement motivation. In C. Ames & R. Ames (Eds.), *Student motivation* (pp. 209–242). New York: Academic Press.

Coursol, A., & Wagner, E. E. (1986). Affect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology, 17,* 136–137.

Cowan, G., & Goldberg, F. J. (1967). Need achievement as a function of the race and sex of figures of selected TAT cards. *Journal of Personality and Social Psychology, 5,* 245–249.

Dahlstrom, W. G., Lachar, D., & Dahlstrom, L. E. (1986). *MMPI patterns of American minorities.* Minneapolis: University of Minnesota Press.

DeBord, L. W. (1977). The achievement syndrome in lower-class boys. *Sociometry, 40,* 190–196.

Drury, D. W. (1980). Black self-esteem and desegregated schools. *Sociology of Education, 53,* 88–103.

Fraser, B. J., Walberg, J., Welch, W. W., & Hattie, J. A. (1987). Synthesis of educational productivity research. *International Journal of Educational Research, 11,* 147–252.

Garza, J. M. (1969). Race, the achievement syndrome, and perception of opportunity. *Phylon, 30,* 338–354.

Graham, S. (1994). Motivation in African Americans. *Review of Educational Research, 64,* 55–117.

Gray-Little, B., & Appelbaum, M. I. (1979). Instrumentality effects in the assessment

of racial differences in self-esteem. *Journal of Personality and Social Psychology, 37,* 1221–1229.

Greene, R. L. (1987). Ethnicity and MMPI performance: A review. *Journal of Consulting and Clinical Psychology, 55,* 497–512.

Greenwald, A. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Gynther, M. D. (1989). MMPI comparisons of Blacks and Whites: A review and commentary. *Journal of Clinical Psychology, 45,* 878–883.

Hall, E. R. (1975). Motivation and achievement in Black and White junior college students. *The Journal of Social Psychology, 97,* 107–113.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88,* 359–369.

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

Hunt, J. G., & Hunt, L. L. (1977). Racial inequality and self-image: Identity maintenance as identity diffusion. *Sociology and Social Research, 61,* 539–559.

Kugle, C. L., Clements, R. O., & Powell, P. M. (1983). Level and stability of self-esteem in relation to academic behavior of second graders. *Journal of Personality and Social Psychology, 44,* 201–207.

Lane, D., & Dunlap, W. (1978). Estimating effect sizes: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107–112.

Lay, R., & Wakstein, J. (1985). Race, academic achievement, and self-concept of ability. *Research in Higher Education, 22,* 43–64.

Leftkowitz, J., & Fraser, A. W. (1980). Assessment of achievement and power motivation of Blacks and Whites, using a Black and White TAT, with Black and White administrators. *Journal of Applied Psychology, 65,* 685–696.

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

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Lott, A. J., & Lott, B. E. (1963). *Negro and white youth: A psychological study in a border-state community.* New York: Holt Rinehart & Winston.

McClelland, L. (1974). Effects of interviewer-respondent race interactions on household interview measures of motivation and intelligence. *Journal of Personality and Social Psychology, 29,* 392–397.

Mingione, A. D. (1965). Need for achievement in Negro and White children. *Journal of Consulting Psychology, 29,* 108–111.

Mingione, A. D. (1968). Need for achievement in Negro, White, and Puerto Rican children. *Journal of Consulting and Clinical Psychology, 32,* 94–95.

Mussen, P. H. (1953). Differences between the TAT responses of Negro and White boys. *Journal of Consulting Psychology, 17,* 373–376.

Nichols, N. J., & McKinney, A. W. (1977). Black or White socioeconomically disadvantaged pupils—they aren't necessarily inferior. *The Journal of Negro Education, 46,* 443–449.

Olsen, H. D. (1972). Effects of changes in academic roles on self-concept of academic ability of black and white compensatory education students. *The Journal of Negro Education, 41,* 365–369.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139–162). New York: Russell Sage Foundation.

Ramirez, M., & Price-Williams, D. R. (1976). Achievement motivation in children of three ethnic groups in the United States. *Journal of Cross-Cultural Psychology, 7,* 49–60.

Rosen, B. C. (1959). Race, ethnicity, and the achievement syndrome. *American Sociological Review, 24,* 47–60.

Rosenberg, M., & Simmons, R. G. (1971). *Black and white self-esteem: The urban school child.* Washington, DC: American Sociological Association

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Ruhland, D., & Feld, S. (1977). The development of achievement motivation in Black and White children. *Child Development, 48,* 1362–1368.

Slavin, R. E. (1986). Best evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15,* 5–11.

Smith, H. P., & Abramson, M. (1962). Racial and family experience correlates of mobility aspiration. *Journal of Negro Education, 31,* 117–124.

Stock, W. A., Okun, M. A., Haring, M. J., Miller, W., & Kinney, C. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11,* 10–14.

Travis, C. B., & Anthony, S. E. (1975). Ethnic composition of schools and achievement motivation. *The Journal of Psychology, 89,* 271–279.

Turner, J. H. (1972). Structural conditions of achievement among Whites and Blacks in the rural South. *Social Problems, 19,* 496–508.

Veroff, J., Atkinson, J. W., Feld, S. C., & Gurin, G. (1960). The use of thematic apperception to assess motivation in a nationwide interview study. *Psychological Monographs, 74,* 1–32.

Veroff, J., & Peele, S. (1969). Initial effects of desegregation on the achievement motivation of Negro elementary school children. *Journal of Social Issues, 25,* 71–91.

Wachter, K. W., & Straf, M. L. (1990). Introduction. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. xiii–xxviii). New York: Russell Sage Foundation.

White, K. R. (1982). The relationship between socioeconomic status and achievement. *Psychological Bulletin, 91,* 461–481.

Wylie, R. C. (1963). Children's estimates of their schoolwork ability, as a function of sex, race, and socioeconomic level. *Journal of Personality, 31,* 203–224.

Wylie, R. C., & Hutchins, E. B. (1967). Schoolwork-ability estimates and aspirations as a function of socioeconomic level, race, and sex. *Psychological Reports, 21,* 781–808.

Authors

HARRIS COOPER is Professor of Psychology, McAlester Hall, University of Missouri, Columbia, MO 65211; crsb0009@mizzou1.missouri.edu. He specializes in meta-analysis and social psychology of education.

NANCY DORR is a graduate student, Department of Psychology, McAlester Hall, University of Missouri, Columbia, MO 65211; c581264@mizzou1.missouri.edu. She specializes in social psychology.

Received October 7, 1994
Revision received December 13, 1994
Accepted May 15, 1995