How Hard Is Hard Science, How Soft Is Soft Science?

The Empirical Cumulativeness of Research

Larry V. Hedges *Department of Education, University of Chicago*

ABSTRACT: Research results in the social and behavioral sciences are often conceded to be less replicable than research results in the physical sciences. However, direct empirical comparisons of the cumulativeness of research in the social and physical sciences have not been made to date. This article notes the parallels between methods used in the quantitative synthesis of research in the social and in the physical sciences. Essentially identical methods are used to test the consistency of research results in physics and in psychology. These methods can be used to compare the consistency of replicated research results in physics and in the social sciences. The methodology is illustrated with 13 exemplary reviews from each domain. The exemplary comparison suggests that the results of physical experiments may not be strikingly more consistent than those of social or behavioral experiments. The data suggest that even the results of physical experiments may not be cumulative in the absolute sense by statistical criteria. It is argued that the study of the actual cumulativeness found in physical data could inform social scientists about what to expect from replicated experiments under good conditions.

Psychologists and other social scientists have often compared their fields to the natural (the "hard") sciences with a tinge of dismay. Those of us in the social and behavioral sciences know intuitively that there is something "softer" and less cumulative about our research results than about those of the physical sciences. It is easy to chronicle the differences between soft and hard sciences that might lead to less cumulative research results in the soft sciences. One such chronicle is provided by Meehl (1978), who listed 20 such differences and went on to argue that reliance on tests of statistical significance also contributes to the poorer cumulativeness of research results in the social sciences. Other distinguished researchers have cited the pervasive presence of interactions (Cronbach, 1975) or historical influences (Gergen, 1973, 1982) as reasons not to expect a cumulative social science. Still others (Kruskal, 1978, 1981) have cited the low quality of data in the social sciences as a barrier to truly cumulative social inquiry. These pessimistic views have been accompanied by a tendency to reconceptualize the philosophy of inquiry into a format that implies less ambitious aspirations for social knowledge (e.g., Cronbach, 1975; Gergen, 1982).

Cumulativeness in the scientific enterprise can mean at least two things. In the broadest sense scientific results are cumulative if empirical laws and theoretical structures build on one another so that later developments extend and unify earlier work. This idea might be called *conceptual* or *theoretical cumulativeness*. The assessment of theoretical cumulativeness must be rather subjective. A narrower and less subjective indicator of cumulativeness is the degree of agreement among replicated experiments or the degree to which related experimental results fit into a simple pattern that makes conceptual sense. This idea might be called *empirical cumulativeness*. The purpose of this article is to suggest that it may be possible to compare at least the empirical cumulativeness of psychological research with that of research in the physical sciences. An exemplary comparison suggests that the differences may be less striking than previously imagined.

The mechanism for this comparison is derived from recent developments in methods for the quantitative synthesis of research in the social sciences. Some of the methods used in meta-analysis are analogous to methods used in the quantitative synthesis of research in the physical sciences. In particular, physicists and psychologists use analogous methods for assessing the consistency of research results, a fact that makes possible comparisons among quantitative reviews in physics and in psychology. One such comparison is reported in this article. This comparison was not chosen in a way that guarantees it to be representative of either social science research or physical science research. However, some effort was exerted to prevent the comparison from obviously favoring one domain or the other, and additional examples are provided to suggest that the case for the empirical cumulativeness of physical science could have been made to look far worse. More data would obviously be needed to support strong conclusions. It seems, however, that the "obvious" conclusion that the results of physical science experiments are more cumulative than those of social science experiments does not have much empirical support.

The Basis for Comparing the Cumulativeness of Research Results in Physical and Social Sciences

It may seem difficult to compare research in the social sciences to research in the physical sciences. Theoretical structures and experimental paradigms are quite different. Each research domain has complications and elaborations that do not arise in the other. Moreover the *meaning* of research results may be quite different. In this article I

Vol. 42, No. 2, 443-455

ignore the many complications and focus instead on aspects of the two domains that can be compared. Experimental results frequently can be expressed as a numerical estimate of a parameter in a theoretical model, such as a mass, an energy, a correlation between variables, or a treatment effect. The consistency of these numerical estimates across replicated experiments *can* be assessed. A comparison of the empirical consistency of the results of replicated experiments in physics (as an example of a physical science) and in psychology (as an example of a social science) is the subject of this article.

Quantitative Research Reviews in the Physical Sciences

Research reviews in the physical sciences serve the same functions as research reviews in the social sciences. Articles that present methodological, conceptual, theoretical, and integrative research reviews can be found in journals such as *Reviews of Modern Physics, Physical Review, Chemical Reviews,* and the *Journal of Physical and Chemical Reference Data.* Integrative research reviews combine evidence from different studies to draw overall conclusions and obtain estimates of parameters. Thus, integrative research reviews in the physical sciences are similar to such research reviews in the social sciences (Cooper, 1984). One difference is that such research reviews in the physical sciences almost always use quantitative methodology to combine research results.

Many research reviews are conducted to establish values for fundamental physical constants. Examples of this type of review are those of Birge (1929), E. R. Cohen (1952), Bearden and Thomsen (1957), and E. R. Cohen and DuMond (1965), who reviewed the literature on various atomic constants. Examples from chemistry are those of Lyman (1961), who reviewed the properties of metals, Ho, Powell, and Liley (1972), who reviewed the thermal properties of elements, and Barton (1975), who reviewed empirical results on solubility parameters. It is also interesting to note that there are several nationally or internationally sponsored efforts to review research in the physical sciences with the aim of providing better data for science and technology. For example, the Committee for Data on Science and Technology (CODATA) of the International Council of Scientific Unions attempts to coordinate the worldwide production of data tables based on critical reviews of available research. The National Standard Reference Data System was created by the United States in the mid-1960s to help serve the same need for compilations of data from critical reviews (Lide & Rossmassler, 1973). The Particle Data Group has been reviewing essentially all of the data reported from research on high-energy physics since 1957 (Rosenfeld, 1975).

Statistical Methods Used in Physical Science Reviews

The most common quantitative method used in reviews in the physical sciences involves the use of weighted least squares (weighted regression). The use of least squares analyses to combine experimental data dates at least from the work of Legendre (1805). The most influential modern proponent of the use of this methodology in reviews of physics experiments was Birge (1932), whose techniques have become standard. Virtually any modern quantitative review of research on physical constants uses Birge's methods. A summary of some of his methodology follows.

Assume that k experiments produce estimates $T_1, \ldots, T_k$ of a parameter θ assumed to have the same value for all experiments. Also assume that the estimate T_i from the i^{th} experiment has a known standard error S_i. Thus, the numerical data available to the reviewer are the estimates $T_1, \ldots, T_k$ and the standard errors $S_1, \ldots, S_k$ from the k experiments.

Birge proposed the use of the weighted average of $T_1, \ldots, T_k$ as an estimate of θ. Specifically, he proposed

$$T_\cdot = \frac{\sum_{i=1}^{k} \omega_i T_i}{\sum_{i=1}^{k} \omega_i}, \tag{1}$$

where $\omega_i = 1/S_i^2$. The weighted average $T_\cdot$ has a standard error given by

$$S(T_\cdot) = [\sum_{i=1}^{k} \omega_i]^{-1/2}. \tag{2}$$

Birge also suggested a method for determining whether the estimates from the k experiments differed by more than unsystematic (measurement) error. He proposed calculating the ratio

$$\mathrm{R} = \frac{\sum_{i=1}^{k} \omega_i (T_i - T_\cdot)^2}{k-1}, \tag{3}$$

which is usually called Birge's ratio. When the results of the experiments are consistent except for sampling error, Birge's ratio is near one. A closely related statistic is

$$X^2 = (k-1)\mathrm{R} = \sum_{i=1}^{k} \omega_i (T_i - T_\cdot)^2, \tag{4}$$

which has a chi-square distribution with $(k-1)$ degrees of freedom when the studies yield consistent results (except for sampling variability). Large values of X^2 (or R) suggest that the results of the k studies disagree.

Birge's procedure for averaging is derived from the use of weighted least squares, and Birge's ratio is the weighted error or residual mean square in that procedure. A generalization of this procedure uses weighted least squares to account for "constrained fits" or expected differences between data from different studies. In this case, the weighted error or residual mean square provides a generalized Birge's ratio, which can be used to determine

The final version of this article was completed while the author was at Michigan State University.

I thank Betsy Jane Becker, David Berliner, Tony Bryk, Nate Gage, Jack Getzels, and Steve Raudenbush for helpful comments on an earlier draft. Kenneth Gergen, Gene Glass, and Paul Meehl revealed their identity as reviewers of this article, and I thank them for their useful comments.

Correspondence concerning this article should be addressed to Larry V. Hedges, Department of Education, The University of Chicago, 5835 S. Kimbark Avenue, Chicago, Illinois 60637.

how well the data from the set of studies agree (except for sampling error). Alternatively, the weighted error or residual *sum* of squares provides a chi-square statistic that is the corresponding generalization of X^2 given in Equation 4. When the data fit the model, this statistic has $(k - p)$ degrees of freedom, where p is the number of constraints on predictors.

It should be noted that physical scientists do not rely blindly on statistical methods. All seem to do a careful qualitative analysis of the research studies, and the data from research studies that seem to have strong sources of bias are omitted from the cumulation (Birge, 1932; Rosenfeld, 1975; Touloukian, 1975). Sometimes studies are omitted simply because they provide estimates that are inconsistent with those from other studies. The Particle Data Group, for example, omits the data from an average of 40% of the available studies from their calculations (Rosenfeld, 1975).

Quantitative Reviews in the Social Sciences

Glass (1976) was among the first authors to recommend the use of quantitative procedures in integrative research reviews in the social sciences. His comprehensive quantitative review of research on the effectiveness of psychotherapy attracted the attention of many psychologists (Smith & Glass, 1977). In the last decade there has been a great deal of interest in the use of statistical methods as a supplement to discursive reviews of research. These quantitative reviews (meta-analyses) typically involve the use of a scale-free index of effect magnitude to express the results of each study. The most popular procedure for combining the results of experimental studies uses the effect size—the standardized difference between the means of the experimental and control groups. Estimates of effect size (sample standardized mean differences) are extracted from each study and are combined across studies to yield an estimate of the average effect size.

Reviewers may also study the covariation of effect sizes with characteristics of the research studies. Such variations are often of interest because theoretical considerations suggest that studies with different characteristics (e.g., using different types of subjects) should yield different results.

Statistical Methods Used in Social Science Reviews

A variety of statistical methods has been used in quantitative reviews in the social sciences. The earliest quantitative research reviews (Glass & Smith, 1979; Smith & Glass, 1977) simply used standard statistical methods to analyze the effect size data. Later researchers (Hedges, 1981, 1982a, 1982b; Kraemer, 1983; Rosenthal & Rubin, 1982a) studied the statistical properties of effect size estimates and showed that the use of standard statistical methods in meta-analysis was suboptimal and sometimes misleading. Several investigators developed alternative statistical procedures designed specifically for meta-analysis. These methods avoid the difficulties that plague the use of conventional statistical methods in research reviews. A summary of some of these methods follows. The summary is not exhaustive because its purpose is to explicate one set of methods that are of particular interest.

Assume that k independent experiments produce effect size estimates $d_1, \cdots d_k$. For the i^{th} experiment

$$d_i = \frac{\bar{X}_i^E - \bar{X}_i^C}{S_i},$$

where $\bar{x}_i^E$ and $\bar{x}_i^C$ are the sample means of the outcome variable in the experimental and control groups and S_i is the pooled within-group sample standard deviation. The effect size estimate d_i estimates the population effect size parameter

$$\delta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i},$$

where μ_i^E and μ_i^C are the *population* means of the outcome variable in the experimental and control groups and σ_i is the population within-group standard deviation in the i^{th} study. I (Hedges, 1981) showed that if the assumptions for the t test between means are met in each study, then the sampling variance of d_i is approximately

$$v_i = \frac{n_i^E + n_i^C}{n_i^E n_i^C} + \frac{d_i^2}{2(n_i^E + n_i^C)}.$$

Therefore $d_1, \ldots, d_k$ are the estimators with standard errors $\sqrt{v_1}, \ldots, \sqrt{v_k}$.

If all of the experiments have the same population effect size δ, that is, if

$$\delta_1 = \cdots = \delta_k = \delta,$$

I (Hedges, 1981) and Rosenthal and Rubin (1982a) proposed the use of a weighted average of $d_1, \cdots d_k$ as an estimator of δ. Specifically, we proposed

$$d_\cdot = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i}, \tag{5}$$

where $w_i = 1/v_i$. The weighted average $d_\cdot$ has a standard error given by

$$S(d_\cdot) = \left[\sum_{i=1}^{k} w_i\right]^{-1/2}. \tag{6}$$

We also proposed a method for determining whether the estimates $d_1, \ldots, d_k$ from the k experiments differed by more than sampling error. We proposed calculating the statistic

$$H = \sum_{i=1}^{k} w_i (d_i - d_\cdot)^2. \tag{7}$$

The homogeneity statistic H has a chi-square distribution with $(k - 1)$ degrees of freedom when $\delta_1 = \cdots = \delta_k$. Of course, large or statistically significant values of H suggest the conclusion that the results of the studies disagree.

I (Hedges, 1982b) noted that these methods for estimating effect size and testing for homogeneity of effect size can also be viewed as weighted least squares procedures. I provided a generalization of these procedures to model expected variation in effect sizes between studies

via a multiple-regression-like linear model. In this case the weighted error or residual sum of squares is a chi-square statistic that quantifies how well the data from a set of studies agree (except for sampling error). When the data fit the model, this chi-square statistic, which I call the "model specification test statistic," has $(k - p)$ degrees of freedom where p is the number of predictors (including the intercept as one predictor).

The Relationship Between Statistical Procedures Used in Physical Science and Social Science Research Reviews

The parallel between the statistical procedures described for physical sciences and those for the social sciences is striking. The calculation of the weighted mean and weighted least squares procedures are completely analogous. Equations 1 and 5 are identical except that d_i and $w_i = 1/v_i$ in Equation 5 replace T_i and $\omega_i = 1/S_i^2$ in Equation 1. Similarly, the chi-square statistics in Equations 4 and 7 are analogous. In fact, both procedures (and analogous procedures for combining correlation coefficients) are a special case of a general result of combining estimates and testing the homogeneity of independent estimators (see, e.g., Rao, 1973, pp. 389–390).

The chi-square statistics test the homogeneity of the parameter estimates by comparing between-experiment and within-experiment estimates of variance. Each term of the form $(T_i - T.)^2$ or $(d_i - d.)^2$ is a kind of variance estimate. In fact, the usual formula for the variance,

$$\frac{1}{k}\sum^{k}(X_i - \bar{X})^2,$$

is merely an average of such terms. Each term of the form $(T_i - T.)^2$ or $(d_i - d.)^2$ is actually an estimate of the variance *between experiments*. Each weight $\omega_i = 1/S_i^2$ or $w_i = 1/v_i$ is the reciprocal of a variance generated from *within* the i^{th} study. Hence each term of the form

$$\omega_i(T_i - T.)^2 = (T_i - T.)^2/S_i^2$$

or

$$w_i(d_i - d.)^2 = (d_i - d.)^2/v_i$$

is actually a ratio of an estimate of variance between experiments to a variance generated within an experiment. If the underlying parameters (θ or δ) are the same across studies, the between- and within-experiment variance estimates should be the same and each term of the sum in the homogeneity statistic should be about one.[1]

The fact that the chi-square statistics for testing the homogeneity of research results are the same suggests that it is possible to compare the homogeneity of research results for groups of experiments in the physical sciences with the homogeneity of research results of groups of experiments in the social sciences. In two groups that have the same number of experiments we can compare the chi-square statistics directly. The larger the chi-square statistic, the greater the heterogeneity of the research results. Alternatively, we can compute Birge's ratio for each group of studies by dividing the chi-square statistic by the number of its degrees of freedom. Under perfect homogeneity among parameters, the expected value of Birge's ratio is one, and larger values indicate more heterogeneity.

The problem of comparing unequal-sized groups of studies is more difficult because the chi-squares (with different degrees of freedom) cannot be compared directly. The expected value of a chi-square equals the number of its degrees of freedom, so chi-squares with greater degrees of freedom are "expected" to be larger. It might seem that Birge's ratio solves the problem, but this is only partly true. The expected value of Birge's ratio will always be one under homogeneity; however, the variance of a chi-square is twice the number of its degrees of freedom, so the Birge's ratios based on a smaller number of studies will be more variable (have a larger variance). Hence a Birge's ratio of 1.5 based on a group of 100 studies is actually far less likely than a Birge's ratio of 2.0 based on a group of 10 studies when both groups of studies have perfectly homogeneous parameters. This suggests the alternative of examining the probabilities (significance levels) associated with the chi-square statistics as an index of homogeneity.

Yet probabilities are also an imperfect means of comparison because the significance levels depend both on the number of studies (the sample size) as well as on the absolute magnitude of the heterogeneity. Thus, reviews with larger numbers of studies will have smaller probability levels than reviews that have the same level of heterogeneity but fewer studies. Because neither Birge's ratio nor the probability level is a perfect index for comparison, both are presented in subsequent analyses. Birge's ratio is emphasized, however, as a direct index of heterogeneity among research results that is independent of sample size.

Some Illustrative Reviews in Physical and Social Sciences

Let us now turn to an analysis of some exemplary reviews. The reviews of physical science research were selected from reviews in physics conducted by the Particle Data Group and reported (Kelly et al., 1980) in *Reviews of Modern Physics*. The reviews of social science research are an eclectic group of reviews in psychology selected to reflect a range of disciplinary specialities and a range of apparent conceptual difficulty in the research area itself.

Illustrative Reviews of Physics

The reviews reported in this section were conducted by the Particle Data Group, an international group of particle physicists headquartered at the University of California, Berkeley. Since 1957, this organization has been collecting published and unpublished experimental results

[1] The conceptual argument given ignores the subtlety that each term $\omega_i(T_i - T.)^2$ or $w_i(d_i - d.)^2$ actually involves the *estimated* mean ($T.$ or $d.$), and consequently the terms are not independent. Taking the dependence of different terms into account, the actual expected value of the sum of the k terms is $k - 1$, the usual degrees of freedom. Thus the expected value of Birge's ratio is one.

on the properties of elementary particles (Rosenfeld, 1975). The Particle Data Group staff evaluates the experimental results and periodically publishes quantitative reviews of this research. The reviews typically take the form of comprehensive reviews of particle properties that attempt to summarize all known properties of elementary particles.

I chose to examine the Particle Group Data reviews for several reasons. First, particle physics is one of the most elite branches of physics. Presumably, many of the very best physicists work in this area. Second, particle physics is an area of great current interest and extraordinary economic investment. This research is expensive—a single experiment can cost $1 million, and new accelerator centers can cost billions of dollars. Third, although some aspects of particle physics are objects of intensive research, many particle properties are well understood, and experimental techniques for measuring these properties are well developed. Finally, the Particle Data Group reviews of particle properties are highly accessible in that they provide complete data on all studies that have been conducted even if the results from those studies are not included in the quantitative estimates of the review.

The consistency of research results varies dramatically with the particle and the particle property under study. The examples presented in this article are reviews of the so-called "stable particles" (stable against strong decay), a class that includes the familiar proton, neutron, and electron. Some stable particles do in fact decay (due to the weak interaction), so most stable particles have a finite lifetime. The example reviews presented here examine studies of the mass and the lifetime of the stable particles. Mass and lifetime were chosen because they are intuitively meaningful properties for which measurement methods are well developed. This choice ensured examination of a research area that is now well understood.

I examined the reviews of mass and lifetime of each stable particle in the Particle Data Group's 1980 review of particle properties (Kelly et al., 1980) and included each case in which the Particle Data Group reported that 10 studies had been completed. Note that this criterion of 10 studies applied to the number of studies conducted, not the number of studies from which data were actually used in the Particle Data Group's quantitative review. In its quantitative review, the Particle Data Group does not use every study that was conducted. All studies are critically evaluated, and roughly 40% of the results are omitted. Results are usually eliminated because (Rosenfeld, 1975) (a) the data reported are preliminary, (b) a standard error is not stated, (c) the data are of poor quality, (d) the result involves assumptions that the Particle Data Group does not wish to incorporate, or (e) the result is inconsistent with other results.

The results of the 13 quantitative reviews that met the criteria are reported in Table 1. The table presents data from both the Particle Data Group's quantitative review and a quantitative review utilizing data from all of the studies that were actually conducted. The latter review corresponds to the common practice of meta-analysts in the social sciences, who attempt to include all of the studies that have actually been conducted. The number of studies, the chi-square fit statistic, and Birge's ratio are also given in each case.

The data presented in Table 1 are striking. When all studies are included in the quantitative reviews, the average Birge ratio is over 2.00, which is 100% larger than

Table 1
Homogeneity Statistics From 13 Particle Data Group (PDG) Reviews

		All studies				Studies in PDG review			
Particle	Property	Number of studies	X^2	p	R	Number of studies	X^2	p	R
Muon	Lifetime	10	29.496	0.000	3.28	9	11.602	0.170	1.45
Charged pion	Mass	10	20.034	0.018	2.23	7	2.046	0.915	0.34
Charged pion	Lifetime	11	34.139	0.000	3.41	10	14.280	0.113	1.59
Neutral pion	Lifetime	11	37.897	0.000	3.79	6	20.205	0.001	4.04
Charged kappa	Lifetime	13	17.633	0.127	1.47	7	14.937	0.021	2.49
Short-lived neutral kappa	Lifetime	13	18.524	0.101	1.54	10	11.415	0.248	1.27
Lambda	Mass	10	39.037	0.000	4.34	5	4.791	0.309	1.20
Lambda	Lifetime	27	70.676	0.000	2.72	3	4.929	0.085	2.46
Sigma +	Lifetime	21	9.834	0.971	0.49	19	8.101	0.977	0.45
Sigma−	Lifetime	16	24.252	0.061	1.62	14	16.808	0.208	1.29
Xi−	Mass	11	9.802	0.458	0.98	9	2.707	0.951	0.34
Xi−	Lifetime	17	11.058	0.806	0.69	11	7.724	0.656	0.77
Omega−	Mass	11	8.611	0.569	0.86	10	8.591	0.476	0.95
M		13.9		0.239	2.11	9.2		0.395	1.43
SD		5.1		0.343	1.28	4.1		0.363	1.05

Note. R is Birge's ratio.

expected when studies yield consistent results. Moreover, 6 of the 13 reviews (46.2%) show statistically significant disagreement among studies.

When only the studies used in the Particle Data Group review are examined, the research results look much more consistent. The average Birge ratio is still nearly 1.5, but only 2 of 13 reviews show statistically significant disagreement among studies. However, the agreement was obtained by deleting 34% of the studies in the reviews, reducing the average number of studies in each quantitative review from 13.9 to 9.2.

Illustrative Reviews in the Social Sciences

Research in the social and behavioral sciences exhibits great diversity. Consequently, the results of quantitative reviews in several different areas of psychology are reported in this section to reflect some of this diversity. I chose these reviews in part either because they used the statistical methodology reported in this article or because the data from each review were available to me.

Five of the reviews were selected from what some might consider a relatively "hard" area of psychology: the study of sex differences in cognitive abilities. Six of the reviews were selected from what some might consider very "soft" areas of educational psychology and evaluation research: studies of the effectiveness of open education programs and studies of the effects of school desegregation on academic achievement. Two of the reviews, which examine studies of the validity of student ratings of instruction and the effect of teacher expectancies on student IQ, are perhaps in a middle ground between very hard and very soft areas of educational psychology. The 13 reviews were obtained from six different publications, each of which is discussed below.

Sex differences in spatial ability. Linn and Peterson (1985) reviewed studies of spatial ability published since Maccoby and Jacklin's (1974) review of psychological gender differences. Linn and Peterson provided a theoretical argument that what is sometimes called spatial ability can actually be divided into three different constructs: spatial perception, mental rotation, and spatial visualization. They then did a quantitative review (meta-analysis) of studies of sex differences on each of these constructs. Because sex differences in spatial ability are usually assumed to emerge in adolescence, Linn and Peterson used a categorical model that grouped results derived from subjects under age 13, subjects aged 13–18, and subjects over 18 years of age. They argued that results within these three groups should be consistent but that between-group differences were the result of the expected emergence of sex differences at adolescence. The categorization was slightly different for the mental rotation tasks. Hence for the purposes of estimating the consistency of research results, I calculated a generalized Birge ratio and chi-square statistic for each of the three reviews reported by Linn and Peterson.

Sex differences in verbal ability and field articulation. Becker and Hedges (1984) reanalyzed the data from a meta-analysis originally published by Hyde (1981). Hyde used quantitative methods to examine studies of sex differences in verbal ability, quantitative ability, visual-spatial ability, and field articulation that were previously reviewed by Maccoby and Jacklin (1974). Becker and Hedges (1984) reanalyzed studies in all four areas, but only the verbal-ability and the field-articulation areas had more than 10 independent results that could be subjected to the desired quantitative analysis. They also used a linear model that included terms to account for bias in the effect size due to differences in sample selectivity and for an expected decrease in sex differences in more recent studies (see Rosenthal & Rubin, 1982b). Hence for the purposes of determining consistency of research results, a generalized Birge ratio and chi-square statistic were calculated for each of the two reviews reported by Becker and Hedges (1984).

The effects of open education on attitude toward school, mathematics achievement, reading achievement, and self-concept. Hedges, Giaconia, and Gage (1981) reviewed the results of randomized experiments on the effects of open education on attitude toward school, mathematics achievement, reading achievement, and self-concept.

The effects of desegregation on educational achievement. Crain and Mahard (1983) reviewed the effects of school desegregation on the academic achievement of black students. The data reported in the present article were derived from a reanalysis of all studies that provided sufficient statistical information to permit using the quantitative methods for testing homogeneity among research results. Crain and Mahard argued that the design of the desegregation study had an effect on study outcomes. Consequently, Crain (personal communication, May 1985) argued that the studies with the strongest designs should be analyzed in two groups: randomized experiments and studies with longitudinal controls. The analysis reported here uses this method.

The validity of student ratings of college faculty. P. A. Cohen (1981) reviewed the literature on validity studies of student ratings of teachers in higher education. The results reported here are based on a reanalysis of Cohen's data on the validity (correlation with achievement) of overall instructor ratings. Note that Cohen's data consisted of correlation coefficients between student ratings and student achievement. The chi-square test and Birge ratio were obtained by procedures analogous to those used for effect sizes.

The effects of teacher expectancy on IQ. Raudenbush (1984) reviewed the literature on randomized experiments of the effects of teacher expectancy on student intelligence. Raudenbush argued that the effect of teacher expectancy is believed to be greater when teachers are not already acquainted with the pupils when the expectancy is induced. Consequently, he grouped the studies according to length of student–teacher contact prior to the inducement of the expectancy. The data reported here represent a slight reanalysis of Raudenbush's data. The data analysis model is designed to estimate a different effect for studies with zero or one week of teacher contact

prior to the inducement of the expectancy from that for studies with more than one week of prior teacher contact.

The results of the 13 reviews are reported in Table 2. The table presents both data from the 13 reviews essentially as analyzed by the authors and data from corresponding reviews in which some studies were deleted as potential "outliers." The reviews as analyzed by the authors exemplified the usual procedure in meta-analyses in the social sciences: No studies were deleted. The reviews with some studies deleted correspond more closely to the practice of the Particle Data Group (and other reviewers in physical science) in which apparently outlying studies are deleted. The number of studies, the *H* statistic, and Birge's ratio are given in each case.

Outliers were deleted by the following procedure. If the homogeneity statistic for all studies was not statistically significant at the $\alpha = .05$ level, no studies were deleted. If the homogeneity statistic for all studies was significant, that study was deleted that yielded the largest reduction in this statistic. Additional studies were deleted using the same criterion until a maximum of 20% of the original studies had been deleted or a high degree of homogeneity was obtained. Note that this procedure for deleting studies actually eliminates substantially fewer studies than does the procedure used in the Particle Data Group reviews.

The data in Table 2 show substantial evidence of disagreement among research studies. When all of the studies are included in the reviews, the average Birge ratio is over 2.00, and 6 of the 13 reviews (46.2%) show statistically significant disagreement among studies.

When there is some deletion of studies, the research results look much more consistent. The average Birge ratio is about 1.3, and none of the 13 reviews shows statistically significant disagreement among studies. However, the agreement was obtained by deleting 4.3% of the studies in the reviews, reducing the average number of studies from 32.1 to 30.8.

A Comparison of the Reviews in the Physical and the Social Sciences

The data reported in Tables 1 and 2 are strikingly similar. When all studies actually conducted are included, reviews in both the physical science and the social science domains suggest statistical inconsistency among research results. In each case the Birge ratio is about 2.0, and there are statistically significant heterogeneities among research results in almost 50% of the reviews. When studies with deviant estimates are deleted, research results in both domains are much more consistent.

The averages of the Birge ratios for the social science reviews are slightly smaller than those of the physical science reviews, indicating that the social science research results are slightly more consistent by this criterion. However, the social science reviews typically involved more studies. Consequently, the average probability value of the chi-square statistics for the social science reviews is also smaller, which indicates that the social science research results are slightly less consistent by this criterion. Neither criterion indicates a very large difference between the consistency of research results from the social sciences and the consistency of those from the physical sciences.

Table 2
Homogeneity Statistics From 13 Social Science Reviews

Review		All studies				Reviews deleting some studies			
		Number of studies	*H*	*p*	R	Number of studies	H	*p*	R
Linn & Peterson (1985)	Spatial perception[a]	62	96.88	0.001	1.64	56	43.24	0.828	0.82
	Spatial visualization[a]	81	98.81	0.056	1.27	81	98.81	0.056	1.27
	Mental rotation[b]	29	43.34	0.024	1.61	29	43.34	0.024	1.61
Becker & Hedges (1984)	Verbal ability[a]	11	32.69	0.000	4.09	9	11.36	0.078	1.89
	Field articulation[a]	14	19.29	0.056	1.75	14	19.29	0.056	1.75
Hedges, Giaconia, & Gage (1981)	Reading achievement	19	105.60	0.000	5.87	16	24.68	0.054	1.65
	Math achievement	17	43.65	0.000	2.73	14	17.40	0.182	1.34
	Attitude to school	11	21.62	0.017	2.16	9	9.00	0.342	1.13
	Self-concept	18	23.66	0.129	1.39	18	23.66	0.129	1.39
Crain & Mahard (1983)	Randomized	13	12.22	0.428	1.02	13	12.22	0.428	1.02
	Longitudinal	57	56.15	0.469	1.00	57	56.15	0.469	1.00
P. A. Cohen (1981)	Validity of student rating	67	104.62	0.002	1.59	65	79.48	0.092	1.24
Raudenbush (1984)	Teacher expectancy[b]	19	17.61	0.414	1.04	19	17.61	0.414	1.04
M		32.1		0.123	2.09	30.8		0.243	1.32
SD		25.0		0.183	1.42	24.8		0.239	0.33

Note. *H* is the chi-square statistic, and R is Birge's ratio or a generalized Birge's ratio.
[a] A three-parameter fit is used. [b] A two-parameter fit is used.

What is surprising is that the research results in the physical sciences are not markedly more consistent than those in the social sciences. The notion that experiments in physics produce strikingly consistent (empirically cumulative) results is simply not supported by the data. Similarly, the notion that experiments in the social sciences produce relatively inconsistent (empirically noncumulative) results is not supported by these data either.

These data do suggest that results from replicated experiments do not always tend to be consistent in an absolute sense (as measured by a statistical test). Almost 50% of the reviews showed statistically significant disagreements in both the social sciences and the physical sciences. The data from the physical sciences show that even research based on sound theories and strong methodology may not always yield results that are consistent in an absolute sense by a statistical criterion. This suggests that caution should be used in any applications of absolute criteria for the consistency of experimental results. Even the best *real* (as opposed to hypothetical) research data may not meet this criterion. If absolute consistency is required, it will often be necessary to delete the results of at least some studies to obtain such consistency. Alternatively, it might be argued that the proper criterion for the consistency of experimental results is relative to the degree of consistency commonly found in areas of research that are generally conceded to produce cumulative research results. This argument would use the degree of consistency commonly found in the physical sciences as a criterion against which to compare research results in the social sciences. The data presented in this article suggest that social science research results are reasonably cumulative by this relative criterion.

Criticisms of the Type of Comparison Presented in This Article

The purpose of this article is not to draw definitive conclusions, but rather to suggest that it might be fruitful to pursue serious comparisons of how research cumulates in the physical as opposed to the social sciences. One possible strategy for such comparisons was presented and illustrated with data from actual research reviews. Although the comparisons presented were intended to suggest that further work in this area could be fruitful, some might contend that they are wildly misleading and should not be taken seriously. In this section I review the most obvious criticisms of the methodology and the illustrative comparisons.

The Reviews From the Physical Sciences Were Not Representative

One potentially serious criticism is that the research domain of the Particle Data Group Reviews, high-energy physics, may not be representative of the physical sciences. Perhaps the measurement of particle properties presents much greater difficulties than other areas of physical science, and therefore the empirical inconsistencies among research studies do not reflect the general state of affairs in physical science. In fact, it is possible that high-energy physics is the only area in the physical sciences to show so much inconsistency among research results.

There is considerable evidence that the empirical inconsistency of research results in particle physics is not unique in the physical sciences. Birge (1932) contended that studies to determine atomic weights yielded results that were very inconsistent. Describing a reanalysis of a major review of atomic weights by Clarke (1920), Birge (1932) found "from sample calculations, that the ratio R_e/R_i [herein described as R, or Birge's ratio] averages about ten" (p. 221). These results are old, but they are clearly far less consistent than those of the Particle Data Group reviews, where the largest value of Birge's ratio for any review was less than five.

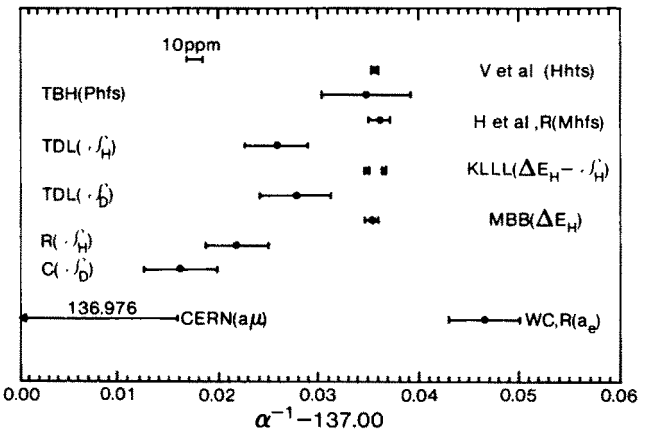
More modern evidence can be found from studies of constants that are important in quantum mechanics. One example of determinations of the fine structure constant α was given in B. N. Taylor, Parker, and Langenberg (1969). The results of 12 experimental determinations of α^{-1} (the inverse of the fine structure constant) are depicted in Figure 1.[2] This figure shows the point estimate of α^{-1} and a one-standard-error interval about that estimate (that is, a 68% confidence interval for α^{-1}). It is clear from the figure that these research results are statistically inconsistent.

Other examples demonstrate significant inconsistencies in measured values of physical constants over time. For example, Bearden and Thomsen (1957) reported the results of several historical reviews of values of the speed of light. They concluded that "it is clear that these values have fluctuated rather drastically, often by several probable errors" (p. 273). E. R. Cohen and DuMond (1965) summarized the history of determinations of Planck's constant, the charge of the electron, the mass of the electron, the fine structure constant, and Avogadro's number from 1929 to 1965. The estimated values of each constant changed significantly over the 36-year period although all of the estimates grew substantially more accurate over time. Even contemporary measurements of presumably well-known constants sometimes disagree significantly. Figure 2 is adapted from Rosenfeld's (1975) plots of other reviewers' estimates of the masses of the proton and electron. The differences are certainly statistically significant. Rosenfeld (1975) concluded "that the reliability is poor" (p. 581).

The determination of chemical and thermodynamic constants provides other examples of substantial inconsistencies among research results. In an article about efforts to critically review data on thermodynamic constants, Zwolinski and Chao (1972) concluded that "the reported values in the literature are usually very incomplete, inconsistent, and at times inaccurate" (p. 115). Another review of reference data on thermodynamics (Touloukian, 1975) provided several striking examples of in-

[2] Figure 1 is adapted from Figure 6 of Taylor, Parker, and Langenberg (1969) by deleting the three data points derived from various reviews of research that were not independent of the results of the individual experiments.

Figure 1
Point Estimate of the Inverse of the Fine Structure Constant and Associated 68% Confidence Intervals From 12 Studies



Note. Adapted from "Determination of e/h, Using Macroscopic Quantum Phase Coherence in Superconductors: Implications for Quantum Electrodynamics and the Fundamental Physical Constants" by B. N. Taylor, W. H. Parker, and D. N. Langenberg, 1969, *Reviews of Modern Physics, 41,* Figure 6, p. 469. Copyright 1969 by the American Physical Society. Adapted by permission.

consistent research results. One particularly illustrative anecdote concerns the serious discrepancies between values that appeared in an important reference work, *The Metals Handbook* (Lyman, 1961), and a systematic redetermination of these values a few years later. In the 1961 edition of this handbook,

> room-temperature thermal conductivity values are given for 64 elements and some other materials. Of the 64 values reported for the 64 elements, 25 are now [as of 1975] known to be in error by over 10%; 16 of the 25 in error by over 30%; 8 of the 16 in error by over 50%; 2 of the 8 in error by over 100%; and one of the two is in error by 245%. (Touloukian, 1975, p. 123)

The Reviews From the Social Sciences Were Not Representative

Another potentially serious criticism is that the illustrative reviews of social science research were not representative of research in psychology or the social sciences. Perhaps the research examined in the illustrative reviews was among the most consistent in the social sciences. In that case, other social science research might show such wild inconsistencies that further comparisons with the physical sciences would be pointless. Such comparisons would only confirm that social science research results are far less consistent than research results in the physical sciences.

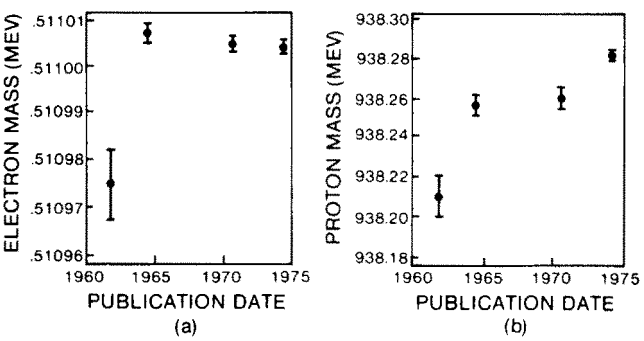
This criticism must be taken seriously. Research in the social sciences is diverse, and the quality of that research is likely to vary widely. The illustrative reviews were chosen from five different areas to at least suggest some of this diversity. Although studies of sex differences in cognitive abilities may present some of the more consistent research results in psychology, it is hard to imagine anyone arguing that research on school desegregation, teacher expectancy, or the effects of open education should yield consistent research findings. The latter three research areas are plagued by problems of inconsistency in the treatment under study. The actual implementation of both open education and school desegregation seems to vary considerably across studies. It is remarkable that the results of these research areas are not even more inconsistent.

The quality of reviews also varies widely. Social scientists have only recently begun to appreciate the need for standards of methodological rigor in research reviews (see Cooper, 1982, 1984; Jackson, 1980). My own experience in reanalyzing research reviews (even quantitative research reviews) is that there are often serious errors in extraction of data from primary studies, in data analysis, and in interpretation. The physical sciences appear to impose more rigorous methodological standards on research reviews. Because I chose the illustrative reviews in social science in part because they used rigorous methodology, they may not be representative of the typical research review.

There is substantial support for the contention that rigorous reviews of some kinds of social science research reveal very consistent results. For example, rigorous reviews of gender differences (Hyde & Linn, 1986) and research on the effectiveness of various methods for teaching composition (Hillocks, 1986) found results similar to those of the reviews examined in this article. Studies of validity generalization (Schmidt & Hunter, 1977) of personnel selection tests have shown that virtually all of the observed variation among the results of validity studies is attributable to sampling error, which implies consistency of research results. In a general treatment of methdology for meta-analysis, Hunter, Schmidt, and Jackson (1982) concluded:

> In our own research in which we have made corrections for sampling error and other artifacts, we have found no significant

Figure 2
Point Estimates of the Mass of the Electron (a) and Proton (b) and Associated 68% Confidence Intervals From Four Reviews Between 1960 and 1974



Note. From "The Particle Data Group: Growth and Operations" by A. H. Rosenfeld, 1975, *Annual Review of Nuclear Science,* Figure 12, p. 582. Copyright 1975 by Annual Reviews, Inc. Used by permission.

remaining variation across studies. That is, it is our experience that there is usually no important variation in study results after sampling error and other artifacts are removed. (p. 32)

Physical Science Measurements Are More Accurate

A different kind of criticism of the comparisons proposed in this article is based on the notion that measurements in the physical sciences are far more accurate. If this is true, then the "real" differences among the outcomes of studies in the physical sciences are trivial in absolute magnitude even if they represent variations of several (infinitesimal) standard errors of estimate. In comparison, the variation among social science research results is larger in absolute magnitude and is therefore more important even if it does not constitute several standard errors of estimate.

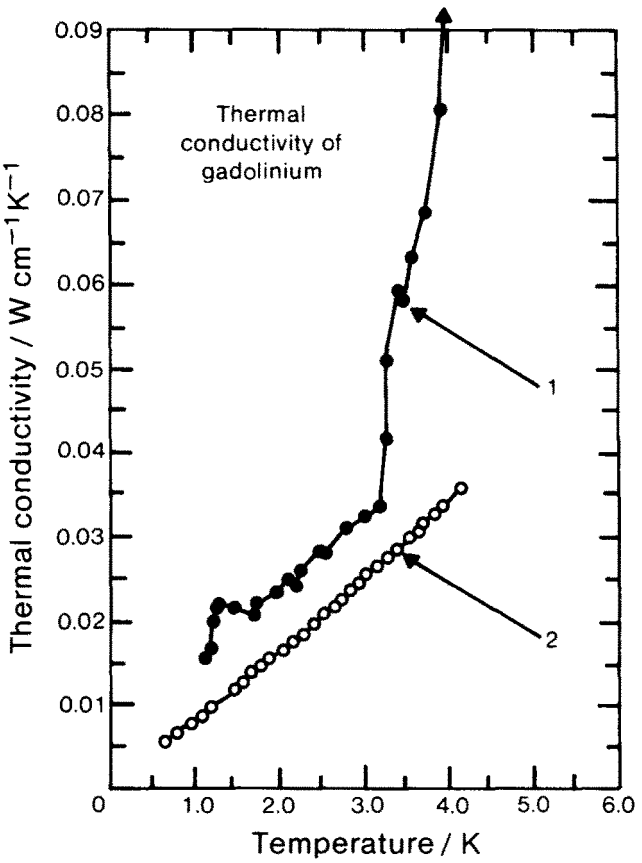
The most immediately obvious difference between measurements in the physical sciences and those in the social sciences does seem to be that measurements in the physical sciences are much more accurate. Results may be quoted in which the uncertainties (standard errors of measurement) are billionths of a second. We often think of the relative accuracy of the measurement as reflected by the ratio of the measured value to its standard error. When the ratio is very large, the measurement is very accurate. Comparing the accuracy of physical measurements with measurements in, for example, reaction-time experiments, it is obvious that the physical measurement is more accurate by several orders of magnitude. Physical measurements are made on true ratio scales, and whenever the outcomes of social or behavioral science studies use measurements on true ratio scales comparisons are straightforward. Such comparisons will almost always favor the physical measurements by a large margin.

Most psychological experiments do not use dependent variables with true ratio scales, however. Instead, psychological experiments usually use cognitive tests or affective measures that are at best on an interval scale of measurement. The question of accuracy of measurement is considerably more complicated in this situation. It is no longer sufficient (and in fact is misleading) to look at the ratio of measured value to its standard error. By a simple transformation (of the form $y = a + bx$), the experimenter can change the ratio of measured value to standard deviation. Indeed, this ratio can be made to equal any desired value. In fact, it is common practice to transform test scores to other metrics to obtain scaled scores that are easier to use. Although it may be argued that some parameters involving differences (e.g., standardized mean differences) do have a well-defined zero point, even these parameters are sometimes transformed for ease of interpretation. For example, the effect size (standardized mean difference) is often transformed (nonlinearly) into a percentile. Thus the ultimate scale that is interpreted need not correspond to the scale of the original parameter even if that scale has a well-defined zero point. The argument here is not that measurements in the social sciences are as accurate as those in the physical sciences; it is that the problem of comparing accuracy is more difficult than it might seem.

It is also important to recognize that there are cases in the physical sciences where the results of experiments are not highly accurate in the sense that the values may differ dramatically from one experiment to another. One example cited by Touloukian (1975) involves two sets of data on the thermal conductivity of gadolinium. The two sets of data (illustrated in Figure 3) "are for the same sample, measured in the same laboratory two years apart in 1967 and 1969. The accuracy of curve 1 was stated as within 1% and that of curve 2 as 0.5% although the curves differ from each other by up to 500% at the higher temperature end" (Touloukian, 1975, p. 123).

Other examples of seemingly large discrepancies between experimental results (implying substantial inaccuracy in at least some of the measurements) are easy to find. The large differences between reported values of thermal conductivities of the chemical elements have already been mentioned. Examples of highly discrepant

Figure 3
Two Sets of Data on the Thermal Conductivity of Gadolinium



Note. From "Reference Data on Thermophysics" by Y. S. Touloukian in *International Review of Science—Physical Chemistry: Vol. 10. Thermochemistry and Thermodynamics* (Figure 4.3, p. 123) by H. A. Skinner (Ed.), 1975, London: Butterworth. Copyright 1975 by Butterworth. Used by permission.

experimental results can be found in other areas than thermochemistry. Data on solubility parameters are often difficult to measure, and many measurements are accurate to only one or two significant figures (Barton, 1983). In astronomy, a major controversy could be resolved if the uncertainty in the measurement of the Hubble constant could be reduced to obtain an accuracy of one significant figure (Nicholl & Segal, 1978). In other areas (such as X-ray crystallography and certain protein assays), the folklore of the research community is that between-laboratory differences are so large that numerical data should only be compared within laboratories. Analytical chemists often perform cooperative studies to try to understand the often large interlaboratory differences in the results of quantitative analyses.

The Constructionist Perspective

The arguments presented in this article depend on the assumption that research findings are strongly a function of underlying empirical laws or processes. Yet there is an increasing tendency to view research findings as (at least partly) constructions of the scientific community that depend on interpretative agreements among members of that community (Feyerabend, 1976; Hanson, 1958; Kuhn, 1962; Lakatos, 1978; Phillips, 1977; Quine, 1960; C. Taylor, 1971). Such interpretive agreements include both decisions about what data are relevant and about what methodological procedures are valid. The principal argument of this article is that the notion of consistency of research results and a quantitative index of consistency are part of the methodological conventions of both the physical and the social sciences. However, to the extent that research findings depend on social constructions, consistency of research results has implications somewhat different from those emphasized in the bulk of this article. From a constructionist point of view, consistency of research results implies either the stability of the social constructions across the contexts in which experiments were conducted or an interpretive norm that leads to the perception of consistency. Thus the most interesting object of study surrounding research results that are perceived to be consistent may be the interpretive agreements that make possible the perception of consistency. One example of such a convention in reviews in physics is the practice of omitting a relatively large proportion of the studies to obtain a consistent sample for data analysis. Similarly, inconsistency among research results might indicate an inconsistency among the interpretative norms of the research community. Once again, the interpretive agreements (or differences therein) that lead to the perception of inconsistency are an important object of study. One example of such a convention in reviews of social science research is the use of statistical hypothesis testing in original research coupled with the use of a methodology that ignores the stochastic properties of such tests in research reviews.

One of the most interesting implications of the constructionist perspective is that the perceived cumulativeness in any research domain is a function of the conventions of evidence and methodology in the research community. Consequently, the study of relative cumulativeness across research domains becomes (at least in part) a study of conventions used by the research community for achieving a sense of cumulativeness. This perspective might provide the starting point for yet other comparative investigations of research in the physical and social sciences.

Conclusions

A fundamental question for any scientific research program is, How cumulative should we expect empirical research results to be? That is, how much consistency should we expect of the results of replicated experiments? Psychologists and other social scientists have often expected more consistency in the outcomes of studies than is possible given the stochastic nature of experimental data. Theoretical analyses of decision strategies used in research reviews have demonstrated the statistical fallacy in, for example, concluding that studies fail to replicate if the outcomes of corresponding significance tests disagree (see Humphreys, 1980, or Hedges & Olkin, 1985, chap. 1). Recent developments in meta-analysis provide more statistically valid means of assessing agreement among the outcomes of replicated experiments. Yet no technical development in statistics can answer the question of what *degree* of agreement should be expected from "good" experimental data based on "good" theory. One way to answer the question of what to expect from "good" scientific research is to examine areas where the quality of research and theory is often thought to be exemplary: the physical sciences. The cumulativeness of replicated experiments in the physical sciences can provide a standard against which to judge the cumulativeness of our results in the behavioral sciences. The degree of cumulativeness of experiments in the physical sciences is probably the most we can expect of behavioral science research. Few social or behavioral scientists would expect our efforts to be *more* cumulative than those in the physical sciences.

The purpose of this article is to suggest that it may be fruitful to compare the empirical cumulativeness of physical and behavioral science data. The evidence presented here suggests that social science research may not be overwhelmingly less cumulative than research in the physical sciences. In fact, the evidence shows several parallels in the reviews of social and physical science domains. Experimental results are not always consistent by statistical criteria. About 45% of the reviews in both domains exhibited statistically significant disagreements when no studies were omitted from the reviews. In both domains the deletion of data from some studies substantially improved the empirical consistency of the research.

The fact that research results in the physical sciences often fail to meet the criterion of statistical consistency has important implications for social and behavioral science. New physical theories are *not* sought on every occasion in which there is a modest failure of experimental consistency. Instead, reasons for the inconsistency are likely to be sought in the methodology of the research

studies. At least tentative confidence in theory stabilizes the situation so that a rather extended series of inconsistent results would be required to force a major reconceptualization. In the social sciences, theory does not often play this stabilizing role.

It should be noted that none of the reviews examined in this article was primarily concerned with directly testing a theory that made specific point predictions. Therefore, these reviews are not suitable for comparing tests of theories in the physical and social sciences. In other situations where the reviews do examine studies that are tests of theories, it is important to recognize that consistency of research results alone does not imply that results are in accord with the theory. Experiments might yield results that are quite consistent but that consistently disconfirm the theory. Thus, statistics that measure the consistency of research results are not an index of the extent to which experimental results conform to predictions based on theory. Indices analogous to Birge's ratio but that measure confirmation of point predictions could be constructed, however.

Finally, note that the data presented in this article do not directly address the issue of conceptual or theoretical cumulativeness. It may well be that the social and behavioral sciences are less theoretically cumulative than the physical sciences. One might also argue that theoretical cumulativeness is really the important issue. Yet empirical cumulativeness is still important. It is difficult to imagine theoretical cumulation without a substantial degree of empirical cumulation. If it is true (as the data presented in this article suggest) that the social and behavioral sciences are not substantially less empirically cumulative than the physical sciences, then any deficiency in theory would not appear to be the result of an inability to generate consistent empirical research results.

REFERENCES

Barton, A. F. M. (1975). Solubility parameters. *Chemical Reviews, 75,* 731–753.

Barton, A. F. M. (1983). *CRC handbook of solubility parameters and other cohesion parameters.* Boca Raton, FL: CRC Press.

Bearden, J. A., & Thomsen, J. S. (1957). A survey of atomic constants. *Nuovo Cimento, Supplement, 5,* 267–360.

Becker, B. J., & Hedges, L. V. (1984). Meta-analysis of cognitive gender differences: A comment on an analysis by Rosenthal and Rubin. *Journal of Educational Psychology, 76,* 583–587.

Birge, R. T. (1929). Probable values of the general physical constants (as of January 1, 1929). *The Physical Review Supplement, 1*(1), 1–73.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review, 40,* 207–227.

Clarke, F. W. (1920). A redetermination of atomic weights. *Memoirs of the National Academy of Science, 16*(3), 1–48.

Cohen, E. R. (1952). The Rydberg constant and the mass of the electron. *Physical Review, 88,* 353–360.

Cohen, E. R., & DuMond, J. W. M. (1965). Our knowledge of the fundamental constants of physics and chemistry in 1965. *Reviews of Modern Physics, 37,* 537–594.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51,* 281–309.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52,* 291–302.

Cooper, H. M. (1984). *The literature review: A systematic approach.* Beverly Hills, CA: Sage.

Crain, R. L., & Mahard, R. E. (1983). The effect of research methodology on desegregation-achievement studies: A meta-analysis. *American Journal of Sociology, 88,* 839–855.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Feyerabend, P. K. (1976). *Against method.* New York: Humanities Press.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology, 26,* 309–320.

Gergen, K. J. (1982). *Toward transformation in social knowledge.* New York: Springer-Verlag.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Glass, G. V, & Smith, M. L. (1979). Meta-analysis of the relationship between class size and achievement. *Educational Evaluation and Policy Analysis, 1,* 2–16.

Hanson, N. R. (1958). *Patterns of discovery.* London: Cambridge University Press.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hedges, L. V. (1982a). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499.

Hedges, L. V. (1982b). Fitting continuous models to effect size data. *Journal of Educational Statistics, 7,* 245–270.

Hedges, L. V., Giaconia, R. M., & Gage, N. L. (1981). *The empirical evidence on the effectiveness of open education.* Stanford, CA: Stanford University School of Education.

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

Hillocks, G. (1986). *Research in written composition: New directions for teaching.* Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills and the National Conference on Research in English.

Ho, C. Y., Powell, R. W., & Liley, P. E. (1972). Thermal conductivity of the elements. *Journal of Physical and Chemical Reference Data, 1,* 279–421.

Humphreys, L. G. (1980). The statistics of failure to replicate: A comment on Buriel's (1978) conclusions. *Journal of Educational Psychology, 72,* 71–75.

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

Hyde, J. S. (1981). How large are cognitive gender differences? *American Psychologist, 36,* 892–901.

Hyde, J. S., & Linn, M. C. (1986). *The psychology of gender: Progress through meta-analysis.* Baltimore, MD: Johns Hopkins University Press.

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438–460.

Kelly, R. L., Horne, C. P., Losty, M. J., Rittenberg, A., Shimada, T., Trippe, T. G., Wohl, C. G., Yost, G. P., Barash-Schmidt, N., Bricman, C., Dionisi, C., Mazzucato, M., Montanet, L., Crawford, R. L., Roos, M., & Armstrong, B. (1980). Review of particle properties. *Reviews of Modern Physics, Supplement, 52.*

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta analysis. *Journal of Education Statistics, 8,* 93–102.

Kruskal, W. (1978). Taking data seriously. In Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray, & H. Zuckerman (Eds.), *Towards a metric of science: The advent of science indicators* (pp. 139–169). New York: Wiley.

Kruskal, W. (1981). Statistics in society: Problems unresolved and unformulated. *Journal of the American Statistical Association, 76,* 505–515.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programmes: Philosophical papers of Imre Lakatos* (Vol 1, pp. 8–101). London: Cambridge University Press.

Legendre, A. M. (1805). On the method of least squares. Translated from the French in D. E. Smith (Ed.), *A source book in mathematics* (pp. 576–579). New York: Dover.

Lide, D. R., & Rossmassler, S. A. (1973). *Annual Review of Physical Chemistry, 24,* 135–158.
Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of gender differences in spatial ability: A meta-analysis. *Child Development, 56,* 1479–1498.
Lyman, T. (1961). *The metals handbook.* Metals Park, OH: American Society for Metals.
Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.
Nicholl, J. F., & Segal, I. E. (1978). Statistical scrutiny of the phenomenological redshift-distance square law. *Annals of Physics, 113,* 1–28.
Phillips, D. L. (1977). *Wittgenstein and scientific knowledge: A sociological perspective.* Totowa, NJ: Rowan & Littlefield.
Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: M.I.T. Press.
Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.
Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76,* 85–97.
Rosenfield, A. H. (1975). The particle data group: Growth and operations. *Annual Review of Nuclear Science,* pp. 555–599.
Rosenthal, R., & Rubin, D. D. (1982a). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504.
Rosenthal, R., & Rubin, D. B. (1982b). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology, 74,* 708–712.
Schmidt, F. L., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.
Smith, M. L., & Glass, G. V (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.
Taylor, B. N., Parker, W. H., & Langenberg, D. N. (1969). Determination of e/h, using macroscopic quantum phase coherence in superconductors: Implications for quantum electrodynamics and the fundamental physical constants. *Reviews of Modern Physics, 41,* 375–496.
Taylor, C. (1971). Interpretation and the sciences of man. *Review of Metaphysics, 25*(1), 3–51.
Touloukian, Y. S. (1975). Reference data on thermophysics. In H. A. Skinner (Ed.), *International review of science physical chemistry: Vol. 10. Thermochemistry and thermodynamics* (pp. 119–146). London: Butterworth.
Zwolinski, B. J., & Chao, J. (1972). Critically evaluated tables of thermodynamic data. In H. A. Skinner (Ed.), *International review of science—physical chemistry: Vol 10. Thermochemistry and thermodynamics* (pp. 93–120). London: Butterworth.