A History of Effect Size Indices

Carl J Huberty

Educational and Psychological Measurement 2002; 62; 227
DOI: 10.1177/0013164402062002002

A HISTORY OF EFFECT SIZE INDICES

CARL J HUBERTY
University of Georgia

Depending on how one interprets what an effect size index is, it may be claimed that its history started around 1940, or about 100 years prior to that. An attempt is made in this article to trace histories of a variety of effect size indices. Effect size bases discussed pertain to (a) relationship, (b) group differences, and (c) group overlap. Multivariable as well as univariate indices are considered in reviewing the histories.

During the past several decades, there has been an exponential increase in the frequency of publications criticizing uses of statistical testing; this pattern has occurred across disciplines as diverse as psychology and wildlife studies (Anderson, Burnham, & Thompson, 2000). Concomitantly, there has been an increased emphasis placed on the reporting and interpretation of effect sizes.

For example, the American Psychological Association (APA) Task Force on Statistical Inference recently emphasized, "Always provide some effect-size estimate when reporting a *p* value" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). The Task Force also wrote,

> Always present effect sizes for primary outcomes. . . . It helps to add brief comments that place these effect sizes in a practical and theoretical context. . . . We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. (p. 599)

The editorial policies of the following 19 journals now require effect size reporting:

- *Career Development Quarterly*
- *Contemporary Educational Psychology*
- *Early Childhood Research Quarterly*
- *Educational and Psychological Measurement*

Educational and Psychological Measurement, Vol. 62 No. 2, April 2002 227-240

- *Exceptional Children*
- *Journal of Agricultural Education*
- *Journal of Applied Psychology*
- *Journal of Community Psychology*
- *Journal of Consulting & Clinical Psychology*
- *Journal of Counseling and Development*
- *Journal of Early Intervention*
- *Journal of Educational and Psychological Consultation*
- *Journal of Experimental Education*
- *Journal of Learning Disabilities*
- *Language Learning*
- *Measurement and Evaluation in Counseling and Development*
- *The Professional Educator*
- *Reading and Writing*
- *Research in the Schools*

It is noteworthy that two of these journals are the flagship journals of the American Counseling Association and the Council of Exceptional Children.

The fifth edition of the APA (2001) *Publication Manual* also emphasizes the importance of effect size reporting:

> For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. You can estimate the magnitude of the effect or the strength of the relationship with a number of common effect size estimates. . . . The general principle to be followed . . . is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (pp. 25-26)

For the past two decades or so, the notion of effect size has been fairly common across introductory statistical methods textbooks, particularly in the behavioral sciences. (An exception to this is the book by Moore, 2000, who is, by background and current position, a bona fide statistician; as good as this book is, it does not include the notion of effect size.) The commonality has not, however, carried over to behavioral science journals that typically report results of quantitative studies, although an increased emphasis on the reporting of effect size index values has taken place very recently. For the past three decades or so, the notion of effect size very commonly pertained to differences between (and sometimes among) means of scores on a single outcome variable. Even to this day, if the expression *effect size* is heard or read, it is estimated that a large percent (> 95%?) of empirical researchers and methodologists will think of the univariate (two-group?) mean-comparison context.

I would maintain that the effect size notion applies to contexts in addition to that involving univariate mean comparisons. These other contexts include, but are not limited to, multiple regression or prediction, multiple correlation,

multivariate analysis of variance, and univariate proportion comparisons. No consistent historical pattern was found across all types of effect size indices. In fact, some indices that are currently considered effect size indices were not originally proposed as effect size indices.

Bases for univariate mean-comparison effect size indices may be categorized according to three interpretations: relationship, group difference, and group overlap. The historical development of these three interpretations of effect size will now be addressed. (As will be evidenced below, these three interpretations may not be considered in any way to reflect "new" thinking.) Some effect size indices in a multiple-response-variable context will follow in a separate section. The article concludes with a Comments section.

Relationship Indices

In the context of data analysis, *relationship* typically refers to the correlation between two characteristics or attributes for a set of analysis units. According to some documentation (e.g., Cowles, 1989, pp. 123, 132; Hald, 1998, p. 164; Johnson & Kotz, 1997, p. 109; Stigler, 1986, p. 298), Francis Galton (1822-1912) originated the concept of correlation in 1889, although a year earlier he used the word *co-relation* (Galton, 1888). Stigler (1999, p. 89) maintained, however, that the concept of correlation was reported some 30 years earlier by Charles Darwin (1809-1882), who was a cousin of Galton. Cowles (1989, p. 141) and Stigler (1986, p. 353) disagreed on whether Auguste Bravis (1811-1863) used the idea of correlation in a 1846 paper. (Stigler, 1986, chap. 9, provides an excellent discussion of the history of [simple] correlation.) It was in 1892 that Francis Y. Edgeworth (1845-1926) used the expression *coefficient of correlation* for the symbol ρ (parameter and statistic were not then commonly differentiated). A disciple of Galton, Karl Pearson (1857-1936), began to popularize the correlation coefficient—he used *r*—around 1896. Some years later, Pearson (1905) defined and labeled η the *correlation ratio*. This coefficient was developed in the context of multiple data arrays (like groups in analysis of variance [ANOVA]) that typically suggested a nonlinear relationship between the grouping variable and the outcome variable. In 1924, Ronald A. Fisher (1890-1962) derived the probability distribution of η in the context of ANOVA. But the explicit analysis connection between the ANOVA *F* test and the correlation ratio was not made until 1935 by Truman L. Kelley (1884-1961). (In an ANOVA context, the η value reflects the correlation between the grouping variable and the outcome variable.) In making the connection, Kelley (1935) proposed an adjustment of the statistic η^2 (to reduce the estimation bias) that he labeled ε^2. (For a more detailed historical discussion of ε^2 and ω^2, along with the respective estimators, see Glass & Hakstian, 1969.) It may be pointed out that the bias in η^2 as an estimator for its population counterpart was recognized by Pearson (1923). (Relationships among η^2, ε^2, and ω^2 and how the latter two reduce but

do not remove the bias related to η^2 are discussed in some detail by Richardson, 1996, pp. 18-19.)

What may be the first textbook that connected η^2 and ε^2 to ANOVA is that by Peters and Van Voorhis (1940, p. 319). It is interesting that the many editions of the two Fisher (1925, 1935) books (*Statistical Methods for Research Workers* and *The Design of Experiments*) did not make the connection. Fisher's lack of attention to the connection was emphasized by Yates (1951) when he mentioned that "research workers . . . pay undue attention to the results of . . . tests of significance . . . and too little to the estimates of the magnitude of the effects they are investigating" (p. 32). (This appears contrary to Kirk's 1996, p. 748, statement that Fisher did make the connection in his 1925 book, *Statistical Methods for Research Workers*.) In a textbook aimed at the behavioral sciences, Diamond (1959) proposed another expression, *differentiation ratio*, for η^2 because "it tells us with what success the groups . . . have been differentiated by the principle which underlies their classification" (p. 55).

A second alternative to η^2 was proposed in a textbook by William L. Hays (1926-1995) in 1963. The Hays (p. 325) index, denoted "est. ω^2," is interpreted as an estimator for the strength of association between a grouping variable and an outcome variable. (Later in his book, Hays, p. 547, used η^2 to denote a *correlation ratio*.) In sum, then, three different strength-of-relationship estimators have been proposed from 1935 to 1963: η, ε, and ω. The proposals for the latter two were made with the idea of reducing estimation bias associated with the first one. (Over the years, notation has been a little bit of an issue in that the three Greek letters are often used to represent sample values.)

When levels of the grouping variable in an ANOVA context are random (rather than fixed), some methodologists suggest that an intraclass correlation coefficient, ρ_I, be used as an effect size index. The random-fixed issue was recognized by Hays (1963, p. 424) and further discussed by Vaughan and Corballis (1969). Richardson (1996, p. 19) provided a more detailed discussion of this issue.

It was alluded to earlier in a two-group research situation that the strength of the relationship pertains to the relationship between the (continuous) outcome variable and the dichotomous grouping variable. If the dichotomy is imposed, then the index of relationship is the biserial correlation coefficient. The biserial *r* idea was suggested by Pearson (1910). According to Stigler (1999, p. 18), Pearson later used the expressions "biserial *r*" and "biserial η." The use of biserial η implies (to me, at least) that the biserial correlation coefficient is simply a special case of η. If the dichotomy is natural (e.g., with gender or with experimental versus control), then the square of the point-biserial correlation coefficient, which is a special case of η^2, could be considered an effect size index, although it was not so considered 90 years ago.

Although indices of relationship have been and are currently considered to assess effect size, it is common to square such an index and consider percent of shared variance to assess the magnitude of an effect. A little history on such a perspective of effect is mentioned earlier in this section. Standard cutoffs for some such indices have been suggested (see, for example, Cohen, 1969, pp. 277-281). Rosenthal and Rubin (1979) pointed out a problem of assessing an r^2 value of, say, .14, as being "small" in some specific research situations. A little later, Rosenthal and Rubin (1982) proposed the binomial effect size display (BESD) as an aid in assessing the "practical importance of any effect indexed by a correlation coefficient" (p. 242).

Finally, an index of relationship that was used with two dichotomous variables was originally proposed by George Udney Yule (1871-1951). Three variations of the Yule (1900) index, *Q*, have been suggested: Pearson coefficient of mean square contingency, Pearson tetrachoric coefficient of correlation, and Tschuprow coefficient (see Cowles, 1989, pp. 142-143, and MacKenzie, 1981, chap. 7, "The Politics of the Contingency Table"). In the current context of comparing two groups using a dichotomous outcome variable, a *Q*-derived value could be used as an effect size value. A popular index that could be used, but seldom is in research practice, is the so-called *C* coefficient named after Harald Cramér (1893-1985), which Cramér (1946, p. 282) actually attributed to Karl Pearson. The Cramér *C* may also be used as an effect size index in the context of comparing multiple proportions. (Use of effect size indices for categorical data was fairly recently discussed by Fleiss, 1994.)

Group Difference Indices

In the two-group mean-comparison situation, the typical effect size index considered is a standardized mean difference. Such an index was proposed by Jacob Cohen (1923-1998) in 1962 when he used the letter *d*. A standardized difference was also included in a discussion by Hays (1963, p. 329), which involved a fairly direct relationship between the population counterpart of *d*, δ, and ω^2. During the 1970s and 1980s, there were some discussions as to which standard deviation should be used as the denominator in *d*. Two suggestions made were the standard deviation pooled across the two groups proposed by Cohen (1969, p. 18) and the standard deviation of the control group—the definition of which is not always clear—proposed by Glass (1976). The letter *d* was used by both Cohen and Glass. (Hedges, 1981, took an exception to these two proposals because of bias in the estimators and suggested an adjusted *d*, denoted by *g*.)

Cohen (1962, p. 148) also proposed a standardized-difference type of an index that might be used in a multiple group context. Here, as in Cohen (1969, p. 267), the letter *f* was used. This index reflects the variability of the group

means relative to a standard deviation. (As pointed out by Cohen, 1969, p. 274, there is a relationship between f and η: $f^2 = \eta^2/(1 - \eta^2)$.) About the same time, Winer (1962, p. 57) proposed an index for "the effect of treatment j" (p. 274): $\tau_j = \mu_j - \mu$, where μ_j is the mean for population j and μ is the grand mean across all of the populations. A short time later, Cohen (1969, p. 269) suggested a standardized mean difference in the context of more than two groups; parameterwise, the index is $\delta = (\mu_{max} - \mu_{min})/\sigma$, where μ_{max} is the largest mean, μ_{min} is the smallest mean, and σ is the standard deviation common to all of the populations involved. (The use of a standardized difference in more complex ANOVA designs is discussed by Olejnik & Algina, 2000, pp. 248-258.)

When the outcome variable is dichotomous, group differences are assessed by comparing two proportions. Cohen (1962) suggested the simple difference in proportions, $|P_1 - P_2|$, as an effect size index in the two-group context. For testing equality of multiple proportions, Cohen (1962) suggested the use of the ratio of the largest proportion to the smallest proportion as an effect size index.

Group Overlap Indices

Building on the earlier work of Kelley (1920, 1923), in 1937, John W. Tilton (1891-1980) suggested that the amount of group overlap be considered—in two-group univariate mean comparisons—in determining whether two means are significantly different. Tilton (1937) proposed that "the comparison of means should be supplemented whenever possible by an explicit measure of overlapping, such as the percentage of area common to the two distributions" (p. 657) and that this calculation be based on "two perfectly normal distributions" (pp. 661-662). Tilton's notion of group overlap as related to two-group statistical testing sat dormant for about 30 years, until it was revisited by Dunnette (1966) and Alf and Abrahams (1968) and a few years later by Elster and Dunnette (1971). Dunnette (1966) restated Tilton's idea: "The greater the amount of overlap, the less effective is the predictor in separating the two distributions" (p. 142). (He used *predictor* for what we currently call an outcome variable.)

Alf and Abrahams (1968) presented a fair bit of detail of calculating the percent of group overlap assuming two normal distributions for the outcome variable. Elster and Dunnette (1971) studied the robustness of Tilton's (1937) measure of overlap when the two distributions of outcome variable scores are nonnormal. Oakes (1986, p. 54) mentioned that the misclassified proportion was considered by Eysenck (1971, p. 34) to distinguish the theoretical interest in differences in IQ scores between races. The concept of group overlap as an effect size basis was also considered by Cohen (1969, pp. 19-21) in the context of a two-group mean comparison. Group overlap was also considered in a two-group context by Kraemer and Andrews (1982) when they suggested

using D as the standard normal deviate that corresponds to the proportion of analysis units in one group that are less than the median score of the other group. (For a recent discussion of some parametric and nonparametric effect size indices, see Hogarty & Kromrey, 2001.)

It was Levy (1967) who may have been the first to relate the notion of group overlap to univariate predictive discriminant analysis (PDA). What he considered was the proportion of misclassified units of analysis into the two groups as a "simple matter to proceed from the usual test of statistical significance to a measure of the substantive significance" (p. 38). (It may be noted that the outcome variable in the original study design will play the role of a predictor variable in the PDA, a conceptual variable role reversal.) The relationship between Levy's idea of group overlap was not explicitly connected to that of Tilton (1937). Some elaboration on the group overlap idea as applied to univariate two-group comparisons was given about 20 years ago by Huberty and Holmes (1983). More recently, Huberty and Lowman (2000) proposed the use of the better-than-chance notion in using group overlap assessed via a PDA as a basis for effect size in the multiple outcome variable context—here, the letter I is used (see also Hess, Olejnik, & Huberty, 2001).

Multivariable Indices

In this section, *multivariable* refers to multiple response variables. A discussion of a design with one or more grouping variables and one response variable was given in the previous sections.

The concept of multiple correlation was originated by Pearson and Lee (1897), and in 1914, Pearson proposed the expression *coefficient of multiple correlation* when he used the symbol R. The association of an effect size index with a multiple regression analysis (MRA) or a multiple correlation analysis (MCA) has been virtually ignored by applied researchers in the behavioral sciences. In relating MRA to ANOVA, Cohen (1977, p. 410) suggested an f-type index, $f^2 = R^2/(1 - R^2)$; the statistical test of interest here is that the true multiple correlation coefficient is zero. (Cohen's use of f here is consistent with what he used in an ANOVA context when $f^2 = \eta^2/(1 - \eta^2)$.) This index reflects a single-to-noise ratio. A better-than-chance effect size index in an MRA or an MCA zero-correlation context was recently suggested by Huberty (1994b): $R^2_{adj} - k/(N-1)$, where R^2_{adj} is an adjusted R^2 value (which depends on whether an MRA or an MCA is the focus), k is the number of X variables, and N is the sample size. The expression $k/(N - 1)$ represents the chance value of R^2 under the null hypothesis that $\rho^2 = 0$; thus, $R^2_{adj} - k/(N-1)$ is a better-than-chance index of effect size. For testing that the true regression weight for X_j is zero, Maxwell (2000, p. 435) suggested using $f^2 = (\rho^2 - \rho^2_{(-j)})/(1 - \rho^2)$ as an effect size index, where $\rho_{(-j)}$ denotes the population multiple correlation coefficient involving all X variables except X_j. (It is not clear if adjusted R^2 values are to be used to calculate Maxwell's sample f^2 value.)

The effect size concept is also applicable in the context of grouping variable effects with multiple outcome variables. This is the multivariate analysis of variance (MANOVA) context. The development and discussion of multivariate indices of strength of relationship appear to have started in the early 1970s. The relevant literature was pretty much summarized by Maurice M. Tatsuoka (1922-1996) in 1973. The use of a multivariate effect size index was first (at least in the behavioral sciences) proposed by Tatsuoka (1970). Tatsuoka (1973, p. 48) and Olejnik and Algina (2000, p. 272) provided other early 1970s references of MANOVA-related effect size indices. As in the univariate mean-comparison context, the proposed MANOVA effect size indices are simple transformations of statistical test criteria. For example, one effect size index is, simply, $\eta^2 = 1 - \Lambda$, where Λ is the MANOVA criterion originated by Samuel S. Wilks (1906-1964) in 1932, which he described as a generalization of the univariate correlation ratio (Cooley & Lohnes, 1971, p. 225). Cramer and Nicewander (1979) proposed three additional indices that may be used as effect size indices in a MANOVA context, one of which is $\tau^2 = 1 - \Lambda^{1/r}$, where $r = \min(p, q)$, p denotes the number of outcome variables, and q denotes the hypothesis degrees of freedom. A little later, Serlin (1982) proposed $\eta^2_{PB} = U / r$, where U denotes the Bartlett-Pillai test criterion. (Transformations of other MANOVA criteria are discussed by Huberty, 1994a, pp. 194-196.)

Just as for the univariate relationship effect size indices, an adjustment was proposed for the multivariate counterparts by Tatsuoka (1973) (see also Huberty, 1994a, p. 195; Olejnik & Algina, 2000; Tatsuoka, 1993).

There is a multivariate index based on group overlap that may be utilized in a group comparison context. Following a one-factor MANOVA, group overlap may be assessed using a PDA. It is recognized that there is a role reversal for the multiple response variables and the lone grouping variable. A PDA hit rate is determined, and then the hit rate is transformed to a better-than-chance index, I, which may serve as an effect size index (Huberty & Lowman, 2000). It should be noted that the I index is applicable under covariance heterogeneity as well as under covariance homogeneity—univariate or multivariate.

An abbreviated time line depicting some originations related to effect size developments is given in Figure 1.

Comments

The recent rise in the popularity of the effect size concept in the behavioral sciences was mentioned earlier in this article. Elmore (2001) recently counted 61 effect size choices. Summaries of many of the available choices have been provided by Cortina and Nouri (2000), Kirk (1996, in press), Rosenthal (1994), Snyder and Lawson (1993), and Thompson (2002).

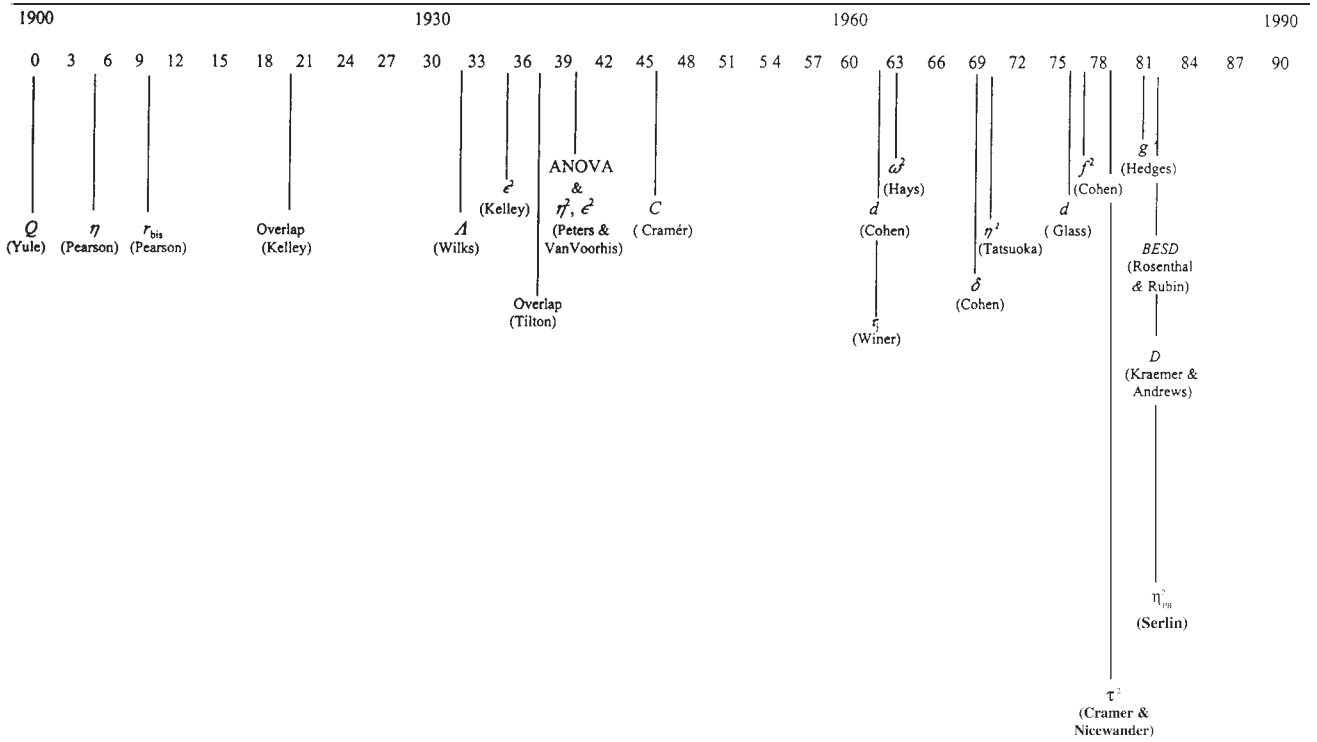

Figure 1. Some approximate years in historical developments related to effect size.

One might hypothesize that this increased popularity of effect size use is a response to the critics of statistical testing. Such criticisms go back a number of decades. There was a large number of concerns pertaining to significance tests and substantive significance in the collection of writings in Morrison and Henkel (1970), but no one explicitly proposed an effect size index. (One of the 31 writings in this book is a critique of statistical testing by Joseph Berkson (1899-1982), M.D., which appeared in a 1942 issue of the *American Statistical Association Journal*.)

How one considers the "effect" in the interpretation process of study results depends, of course, on the study purpose and on the study design. There are study purposes that pertain to intervariable relationships, group mean differences, and group proportion differences. With respect to a comparative-group design, the effect size index to consider in a multiple-group context may depend on whether the groups are independent (between-groups design), dependent (within-groups design), or both (split-plot design) and on the number of outcome variables (see Olejnik & Algina, 2000, for elaborations; see also Fern & Monroe, 1996, for a discussion of related restrictions in the use of effect size indices).

Numerical indices that are now currently used or could be used to reflect a magnitude of effect have been available for a number of decades. As one might surmise, some relationship indices are those originated for a purpose other than to reflect size of effect in a group-comparison study, for example, η^2.

A number of limitations to the use of effect size indices in comparative studies were pointed out by Olejnik and Algina (2000), four of which are

- Limited reliability of outcome variable scores
- Outcome variable variance heterogeneity
- Design quality
- Definition of grouping variable levels

How, in general, is an effect size index value utilized? Of course, it is utilized in the context of statistical testing wherein the researcher arrives at a referent distribution tail-area value, a probability value denoted here by *P*. Suppose the researcher also determines an effect size index value, say, *E*. Two approaches to using the *P* and *E* values are the following:

1. Using *P*, decide if an effect is obtained, and then use *E* to determine how big the effect is.
2. Consider the *P* value and the *E* value jointly; if the *P* value is small and the *E* value is substantial, then a real effect is obtained.

The predominance of statistical testing in the behavioral sciences, at least, has led to some standards pertaining to magnitudes of *P*. The dominant use of

$P = .05$ as a standard across all types of research studies and across all types of statistical analyses is somewhat puzzling. Just as puzzling is the use of some cutoffs for describing magnitudes of E. The interpretation of the index value magnitude is, perhaps, the biggest limitation of the use of E. It appears that the only cutoffs to which applied researchers have paid attention are those standards initiated by Cohen (1969). As astutely noted by Olejnik and Algina (2000), "There is little empirical justification for these standards" (p. 277). Furthermore, as Thompson (2001) recently noted regarding Cohen's criteria, "If people interpreted effect sizes with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (pp. 82-83). As admirable as it was for Cohen (1969) to initiate some effect size magnitude guidelines, much more empirical and design research is needed to establish guidelines for different designs and different data conditions. With respect to the latter, very little thought has been given to the common condition of unequal variances or unequal covariance matrices in group comparison studies.

There is an approach to effect size estimation that may be useful in a group comparison context with one or more outcome variables and with or without variance/covariance homogeneity. The index proposed is based on group overlap and involves classifying analysis units into the criterion groups. (More thinking is needed for multiple grouping variables.) The classification accuracy may be transformed to an index, I, that is a better-than-chance classification index. Some very initial guidelines for classification accuracy were proposed by Huberty and Holmes (1983) and for I values by Huberty and Lowman (2000).

References

Alf, E., & Abrahams, N. M. (1968). Relationship between percent overlap and measures of correlation. *Educational and Psychological Measurement*, *28*, 779-792.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, *64*, 912-923.

Berkson, J. (1942). Tests of significance considered as evidence. *American Statistical Association Journal*, *33*, 325-335.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cooley, W. W., & Lohnes, P. R. (1971). *Multivariate data analysis*. New York: Wiley.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.

Cowles, M. (1989). *Statistics in psychology: An historical perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.
Cramer, E. M., & Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika*, *44*, 43-54.
Diamond, S. (1959). *Information and error.* New York: Basic Books.
Dunnette, M. D. (1966). *Personnel selection and placement*. Belmont, CA: Wadsworth.
Edgeworth, F. W. (1892). Correlated averages. *Philosophical Magazine* (5th series), *34*, 190-204.
Elmore, F. (2001, April). *A primer on basic effect size concepts*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
Elster, R. S., & Dunnette, M. D. (1971). The robustness of Tilton's measure of overlap. *Educational and Psychological Measurement*, *31*, 685-697.
Eysenck, H. J. (1971). *Race, intelligence and education*. London: Temple Smith.
Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*, 80-105.
Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. *Proceedings of the International Congress of Mathematics*, *2*, 805-813.
Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.
Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.
Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York: Russell Sage Foundation.
Galton, F. (1988). Co-relations and their measurement. *Proceedings of the Royal Society of London*, *45*, 135-145.
Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.
Glass, G. V, & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, *6*, 403-414.
Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. New York: Wiley.
Hays, W. L. (1963). *Statistics for psychologists*. New York: Holt, Rinehart & Winston.
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.
Hess, B., Olejnik, S., & Huberty, C. J (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement*, *61*, 909-936.
Hogarty, K. Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can we guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
Huberty, C. J (1994a). *Applied discriminant analysis*. New York: Wiley.
Huberty, C. J (1994b). A note on interpreting an R^2 value. *Journal of Educational and Behavioral Statistics*, *19*, 351-356.
Huberty, C. J, & Holmes, S. E. (1983). Two-group comparisons and univariate classification. *Educational and Psychological Measurement*, *43*, 15-26.
Huberty, C. J, & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, *60*, 543-563.
Johnson, N. L., & Kotz, S. (Eds.). (1997). *Leading personalities in statistical sciences*. New York: Wiley.
Kelley, T. L. (1920). Measurement of overlapping. *Journal of Educational Psychology*, *11*, 458-461.
Kelley, T. L. (1923). *Statistical method*. New York: Macmillan.
Kelley, T. L. (1935). An unbiased correlation ratio. *Proceedings of the National Academy of Sciences*, *21*, 554-559.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.
Kirk, R. E. (in press). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology.* Oxford, UK: Blackwell.
Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, *91*, 404-412.
Levy, P. (1967). Substantive significance of significant differences between two groups. *Psychological Bulletin*, *67*, 37-40.
MacKenzie, D. A. (1981). *Statistics in Britain, 1865-1930*. Edinburgh, UK: Edinburgh University Press.
Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, *5*, 434-458.
Moore, D. S. (2000). *The basic practice of statistics*. New York: Freeman.
Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.
Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chichester, UK: Wiley.
Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241-286.
Pearson, K. (1905). Mathematical contributions to the theory of evolution, XIV: On the general theory of skew correlation and non-linear regression (*Drapers' Company Research Memoirs*, Biometric Series II). London: Dulau.
Pearson, K. (1910). On a new method of determining correlation, when one variable is given by alternative and the other by multiple categories. *Biometrika*, *7*, 248-257.
Pearson, K. (1914). On certain errors with regard to multiple correlation occasionally made by those who have not adequately studied this subject. *Biometrika*, *10*, 181-187.
Pearson, K. (1923). On the correction necessary for the correlation ratio, η. *Biometrika*, *14*, 412-417.
Pearson, K., & Lee, A. (1897). On the distribution of frequency (variation and correlation) of the barometric height of divers stations. *Philosophical Transactions of the Royal Society of London*, *190*, 423-469.
Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: McGraw-Hill.
Richardson, J.T.E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers*, *28*, 12-22.
Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, *9*, 395-396.
Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166-169.
Serlin, R. C. (1982). A multivariate measure of association based on Pillai-Bartlett procedure. *Psychological Bulletin*, *91*, 413-417.
Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, *61*, 334-349.
Stigler, S. M. (1986). *The history of statistics*. Cambridge, MA: Belknap.
Stigler, S. M. (1999). *Statistics on the table*. Cambridge, MA: Harvard University Press.
Tatsuoka, M. M. (1970). *Discriminant analysis: The study of group differences*. Champaign, IL: Institute for Personality and Ability Testing.
Tatsuoka, M. M. (1973). *An examination of the statistical properties of a multivariate measure of strength of association*. Final Report to U.S. Office of Education on Contract No. OEG-5-72-0027.

Tatsuoka, M. M. (1993). Effect size. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 461-479). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80-93.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Tilton, J. W. (1937). The measurement of overlapping. *Journal of Educational Psychology*, *28*, 656-662.

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimated strength of effects in selected ANOVA designs. *Psychological Bulletin*, *72*, 204-213.

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604. [Reprint available from http://apa.org/journals/amp/amp548594.html]

Wilks, S. S. (1932). Certain generalizations of the analysis of variance. *Biometrika*, *39*, 471-494.

Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *American Statistical Association Journal*, *46*, 19-34.

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society, A*, *194*, 257-319.