Psychological Methods
2005, Vol. 10, No. 2, 178–192

1082-989X/05/$12.00 DOI: 10.1037/1082-989X.10.2.178

Use of the Extreme Groups Approach: A Critical Reexamination and New Recommendations

Kristopher J. Preacher
University of North Carolina at Chapel Hill

Derek D. Rucker
Ohio State University

Robert C. MacCallum
University of North Carolina at Chapel Hill

W. Alan Nicewander
Pacific Metrics Corporation

Analysis of continuous variables sometimes proceeds by selecting individuals on the basis of extreme scores of a sample distribution and submitting only those extreme scores to further analysis. This sampling method is known as the extreme groups approach (EGA). EGA is often used to achieve greater statistical power in subsequent hypothesis tests. However, there are several largely unrecognized costs associated with EGA that must be considered. The authors illustrate the effects EGA can have on power, standardized effect size, reliability, model specification, and the interpretability of results. Finally, the authors discuss alternative procedures, as well as possible legitimate uses of EGA. The authors urge researchers, editors, reviewers, and consumers to carefully assess the extent to which EGA is an appropriate tool in their own research and in that of others.

Keywords: extreme groups, tertile split, quartile split, dichotomization

In psychological research, one technique used to examine the relationship between two variables *x* and *y*, when at least *x* is continuous, consists of selecting individuals on the basis of extreme scores of *x* (most commonly, upper and lower tertiles or quartiles) and examining the relationship between *x* and *y* only for those extreme scoring individuals. For example, Pontari and Schlenker (2000) assessed accuracy of recall in a role-playing task only for people scoring in the outer tertiles of an introversion/extraversion scale. Similarly, Deffenbacher, Huff, Lynch, Oetting, and Salvatore (2000) administered the Driving Anger Scale to a large sample as a screening instrument, retaining for further analysis only those individuals scoring in the upper and lower quartiles. Likewise, a researcher interested in the effects of extraversion on social interaction might use a quartile split to identify individuals who are "clearly" extraverted versus introverted and then examine the behavior of only those individuals in conversations, with the intent of generalizing results to individuals across the full range of extraversion. We refer to this and similar sampling procedures collectively as the *extreme groups approach* (EGA). For additional recent examples see Bernichon, Cook, and Brown (2003); Cross, Morris, and Gore (2002); and Verplanken and Holland (2002).

Various modifications of EGA exist. For example, extreme groups need not be equal in size or cover the same range of scores. For a scale for which scores range from 2.1 to 11.6, the low group may have scores ranging from 2.1 to 2.7, whereas the high group may have scores ranging from 7.8 to 11.6. Extreme groups are sometimes chosen on the basis of sample-dependent quantiles, cutoff points derived from population norms, or even the inherent nature of the scale itself. Finally, the scores retained as a result of EGA are often coded and analyzed in terms of low versus high (or young versus old, etc.), reducing individual differences to a simple binary code. This practice involves ignoring individual-differences information in favor of creating quasi-arbitrary groups assumed to be homogeneous on the variable of interest. We wish to emphasize that EGA, as we define it, does not encompass studies that limit generalizations to one extreme or the other, such as those restricting attention to severely depressed people or to academically gifted children.

Kristopher J. Preacher and Robert C. MacCallum, Department of Psychology, University of North Carolina at Chapel Hill; Derek D. Rucker, Department of Psychology, Ohio State University; W. Alan Nicewander, Pacific Metrics Corporation, Monterey, California.

This work was funded in part by National Institute on Drug Abuse Grant DA16883, awarded to Kristopher J. Preacher. We thank Michael C. Edwards and Charles M. Judd for helpful comments.

Correspondence concerning this article should be addressed to Kristopher J. Preacher, CB#3270 Davie Hall, University of North Carolina, Chapel Hill, NC 27599-3270. E-mail: preacher@unc.edu or maccallum@unc.edu

EGA has been widely used in various disciplines for several decades. An early use of EGA is found in a study published by Alfred Binet (1900). Binet used a sample of 11 children selected as being the 5 most intelligent (*intelligents*) and the 6 least intelligent (*inintelligents*) out of a class of 32 children. The performance of these children on a large battery of mental tests (e.g., reaction time, digit memory, counting) was examined with respect to the children's presumed intellectual capacity. EGA was considered in a series of methodological papers published mainly in the 1940s and 1950s. Kelley (1939) proposed sampling extreme groups to enhance statistical power. Some methodologists (Bartlett, 1949; Gibson & Jowett, 1957; Nair & Banerjee, 1942; Nair & Shrivastava, 1942; Wald, 1940) proposed creating subgroups to simplify the complex task of fitting a line to data. Peters (1941; Peters & Van Voorhis, 1940) advocated the use of extreme groups for the purpose of cutting costs associated with data collection. Subsequent to these publications, EGA has been used in numerous examples of applied research throughout the social sciences for a variety of reasons, and its popularity has not waned in recent years. This is evident by the broad use of variants of EGA in articles appearing in numerous top-ranked American Psychological Association journals between 1999 and 2004, including but not limited to, *Behavioral Neuroscience, Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Comparative Psychology, Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology,* and *Journal of Personality and Social Psychology*. In addition, EGA is used with some regularity in the field of genetics (e.g., for selective genotyping; Darvasi & Soller, 1992; Henshall & Goddard, 1999; Muranty & Goffinet, 1997).

Despite the use of EGA across a wide variety of disciplines within the social sciences, a thorough, modern consideration of its advantages and disadvantages is lacking. Several authors have investigated statistical advantages granted by EGA (e.g., Abrahams & Alf, 1978; Alf & Abrahams, 1975; Borich & Godbout, 1974; Feldt, 1961; Flanagan, 1939; Kelley, 1939). Furthermore, several have voiced support for its use (e.g., Kagan, Snidman, & Arcus, 1998; Sorrentino & Short, 1977; Torgesen, 1991). However, EGA has also been the target of much criticism. Humphreys and Dachler (1969a, 1969b) and Humphreys (1978) addressed drawbacks associated with EGA followed by analysis of variance (ANOVA). McClelland and Judd (1993) pointed out that it is unwise to discard data when testing for interaction effects. Wherry (1984) addressed some statistical consequences of EGA, calling it "a favorite dodge of lazy researchers" (p. 49).

Our goal is to inform researchers of both the benefits and costs associated with EGA. We provide a summary of past discussions and critical analyses of EGA, as well as a more comprehensive investigation of both the benefits and costs associated with this technique. In doing so, we provide new recommendations about when the procedure is likely to be appropriate versus inappropriate, as well as a discussion of precautions that should be taken when analyzing extreme groups data.

Past Statistical Investigations of EGA

Feldt (1961) presented one of the first serious examinations of EGA. In Feldt's treatment, measures x and y are assumed to bear some substantively interesting linear relationship and to be bivariate normally distributed in the population. The purpose of Feldt's investigation was to determine the percentage of the sample that must be included in the tails of the x distribution in order to maximize the statistical power of the t test of the difference between x group means on variable y. Feldt found that, assuming normality, maximum power is achieved when the proportion included in each tail is somewhere between .25 and .27 (i.e., approximately a quartile split) and that this proportion remains remarkably constant over a wide range of population correlations.[1] Feldt further investigated the difference in power between the traditional correlational approach, in which all values for x and y are retained, and a t test conducted after a quartile split on x, assuming varying levels of population correlation between x and y. He found that a t test following EGA often provided a more powerful test, unless retaining more than about 80% of the sample was feasible, in which case he recommended using a correlational approach instead.

Alf and Abrahams (1975) extended Feldt's (1961) work by comparing three analytic strategies for examining an x–y relationship. Strategy 1 involved selecting a subsample on the basis of extreme x scores and correlating y with the

[1] An identical finding was reported earlier by Jensen (1928; derivation due to T. L. Kelley) in the context of job placement in education and by Kelley (1939) in the context of item validation. When x and y are bivariate normally distributed, when the retained x scores are dichotomized, and when $\rho_{xy} = 0$, the variance of the maximum-likelihood estimate of ρ_{xy} is minimized when the proportion retained in each tail of the x distribution is $\lambda = .2702$. This finding, which came to be known as the *twenty-seven per cent rule*, was later supported analytically by Mosteller (1946), Kelley (1947), Cureton (1957), and McCabe (1980) and empirically by Garg (1983) via simulation. Flanagan (1939) used this rule to create a chart to ease the burden of computing Pearson product–moment correlations in the context of item validation. Feldt (1961) found that as ρ_{xy} increases, λ decreases but stays in the neighborhood of .27. D'Agostino and Cureton (1975) showed that as the correlation between x and y increases, the optimal proportion of cases that should be retained in each tail to maximize power of the subsequent t test approaches .21. Contrary to these findings, Ross and Weitzman (1964) and Ross and Lumsden (1964) showed that as ρ_{xy} increases, λ also increases. Both camps showed that the power to reject the null hypothesis $\rho_{xy} = 0$ is maximized when λ is in the neighborhood of .27 and ρ_{xy} is small.

extreme x scores (Strategy 1 is equivalent to EGA). Strategy 2 involved selecting a subsample on the basis of extreme x scores and comparing group means on y, an approach identical to that of Feldt. Strategy 3 involved simply retaining the full range of continuous scores on x and y and computing the sample correlation r_{xy}. Given equal sample sizes for all three strategies, after investigating the power of each approach, the authors concluded that Strategy 1 was universally more powerful than Strategies 2 or 3, except when the Strategy 1 data were composed of scores from the full range of x. Under the latter condition, Strategies 1 and 3 are formally equivalent, and both were found to be superior to Strategy 2.

Uses and Misuses of EGA

In the following sections we examine in detail some of the more common rationales offered for the selection and analysis of extreme groups, including increased cost-efficiency, improved power, enhanced effect size, and improved reliability. We explore the legitimacy of these rationales and discuss potential pitfalls. In subsequent sections we describe some potentially legitimate applications of EGA and offer some recommendations for practice.

EGA and Cost-Efficiency

Cost can be a limiting factor in data collection. A measure of some variable y may be expensive to administer or may require so much time to administer that obtaining a large sample is not feasible. EGA was originally developed largely to reduce the sample size necessary to observe an effect without compromising statistical power (Abrahams & Alf, 1978; Alf & Abrahams, 1975; Feldt, 1961; Peters, 1941). Given a fixed sample size, EGA improves cost-efficiency by allowing researchers to selectively sample those regions of the x distribution that will maximize the power of subsequent tests. Improvement of cost-efficiency is a clear benefit of EGA.

EGA and Statistical Power

Perhaps the primary reason for the continued use of EGA is the belief that it increases power in subsequent statistical tests (see, e.g., the advice of A. Tybout in Iacobucci, 2001, pp. 48–49). This assumption is correct and is straightforward to demonstrate. We now illustrate the effect of EGA on power both analytically and via simulation. Alf and Abrahams (1975) presented a power parameter (noncentrality parameter) for t tests of correlations, which they borrowed and adapted from Feldt (1961), who in turn borrowed and adapted it from E. S. Pearson and Hartley (1956):

$$\phi_r = \frac{\rho_{xy}\sqrt{pN}}{\sqrt{1-\rho_{xy}^2}}, \qquad (1)$$

where ρ_{xy} represents the population correlation using full-range data, N is the original full-range sample size, and p represents the proportion of N retained in each half of the distribution of x, assuming normality. Thus, $2p$ is the proportion of N used in the analysis. They presented a similar noncentrality parameter for tests of correlations computed with extreme groups:

$$\phi_{r'} = \frac{\rho_{xy}\sqrt{\left(1+\frac{xh}{p}\right)pN}}{\sqrt{1-\rho_{xy}^2}}, \qquad (2)$$

where x is the z-score cutting off the top pN subjects, and h is the ordinate of the standard normal distribution at x. Thus, both ϕ_r and $\phi_{r'}$ involve using the same sample size, $2pN$, and Equation 2 simplifies to Equation 1 when EGA is not used, that is, when $x = 0$. Given x, h is easily calculated as follows:

$$h = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)}. \qquad (3)$$

The xh/p term is never negative and increases with the extremity of group selection. Therefore $\phi_{r'}$ will be larger than ϕ_r for more extreme groups, illustrating that all else being equal, EGA tends to increase power, with larger increases resulting from more extreme selection.

Operating on a suggestion of Alf and Abrahams (1975), we present a short illustration of the effects of EGA on power. The quantity pN was held constant at 50 so that the curtailed sample size was a constant 100. For full sample $N = 100–500$, the p necessary to yield $2pN = 100$ was calculated. The z scores (x) corresponding to those values of p were then computed. Alf and Abrahams reported the ratio of $\phi_{r'}$ to ϕ_r with the following expression:

$$\frac{\phi_{r'}}{\phi_r} = \sqrt{1+\frac{hz}{p}}. \qquad (4)$$

The ratio of noncentrality parameters conveniently does not depend on either ρ_{xy} or N. The ratio $\phi_{r'}/\phi_r$ thus permits indirect examination of the relative power of tests of r' and r for any values of N and ρ_{xy}. For $p = \{0.001\ldots 0.500\}$, in steps of .001, the ratio $\phi_{r'}/\phi_r$ was calculated[2] and plotted, with the results shown in Figure 1. Two points worth noting are that (a) the relation is never less than 1.0, indicating that, apart from sampling variability, power is generally enhanced after EGA relative to no extreme group selection, and (b) the more extreme the extreme groups become (i.e., the smaller p is), the greater the power benefits become.

To illustrate the effects of EGA on power more directly,

[2] x depends on p, and h depends on x, so p is all that needs to be manipulated.

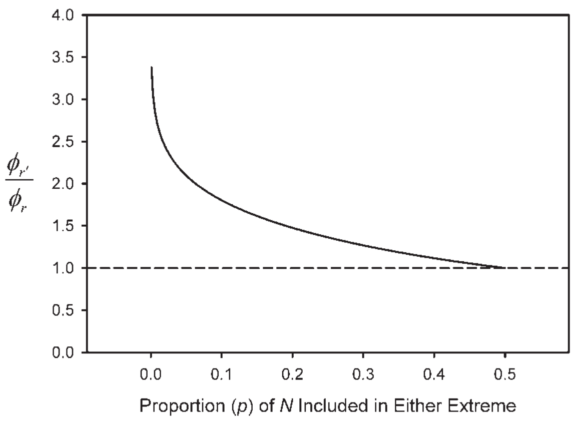

Figure 1. Ratio of the noncentrality parameters ($\phi_{r'}/\phi_r$) for correlation coefficients computed by use of curtailed (extreme groups) data and full-range data under conditions of the extreme groups approach (EGA). Noncentrality is a measure of effect size in units of a test statistic. This ratio is never less than 1.0, meaning that power is always enhanced by EGA, with greater power resulting from more extreme splits.

we generated data corresponding to cells of a 3 (ρ_{xy} = .1, .3, .5) × 6 (N = 24, 60, 120, 240, 360, and 480) × 3 (proportion of data omitted = .00, .33, .50) design. The proportions of data omitted were chosen to reflect no data omitted, a tertile split, and a quartile split, respectively. These splits were selected because they represent common types of EGA found in the literature. For each cell, 10,000 samples of raw data were generated with a procedure described by Kaiser and Dickman (1962). Each sample was generated to be of size $N(1 - p_o)^{-1}$, where p_{o} is the proportion of data omitted. For each sample, two variables were generated and their sample correlation (r_{xy}) was computed. The statistical significance of r_{xy} was determined by calculating the t statistic associated with r_{xy}:

$$t = \frac{r_{xy}\sqrt{N - 2}}{\sqrt{1 - r_{xy}^2}} . \qquad (5)$$

Results are reported in Figure 2. Each panel in Figure 2 corresponds to a value of ρ_{xy}. Each point in the plots represents the proportion of times, out of 10,000, in which t exceeded the critical value for one-tailed significance at α = .05 and $df = N - 2$. This proportion can be interpreted as empirical power. As expected, power increases as N and ρ_{xy} increase. It was also found (although not depicted in Figure 2) that the proportion of significant results fluctuates around α when $\rho_{xy} = 0$ and that power was uniformly close to 1.0 for values of ρ_{xy} above .5. Comparison of lines within each panel shows a clear tendency for successively more extreme splits to more strongly enhance empirical power, especially for small sample sizes.

In summary, EGA will usually improve power relative to analysis of full-range data. However, increases in power are not necessarily always desirable. We reiterate the frequent admonition (e.g., Kirk, 1996, 2001; Wilkinson & the APA Task Force on Statistical Inference, 1999) that the primary focus of research should not be to obtain significant p values but rather to determine what the data tell us about the phenomena of interest—that is, effect size and practical significance. Judging EGA to be appropriate because it increases the power of a subsequent statistical test represents a focus on significance-seeking. Whereas EGA may be applied for the purpose of making efficient use of a sample

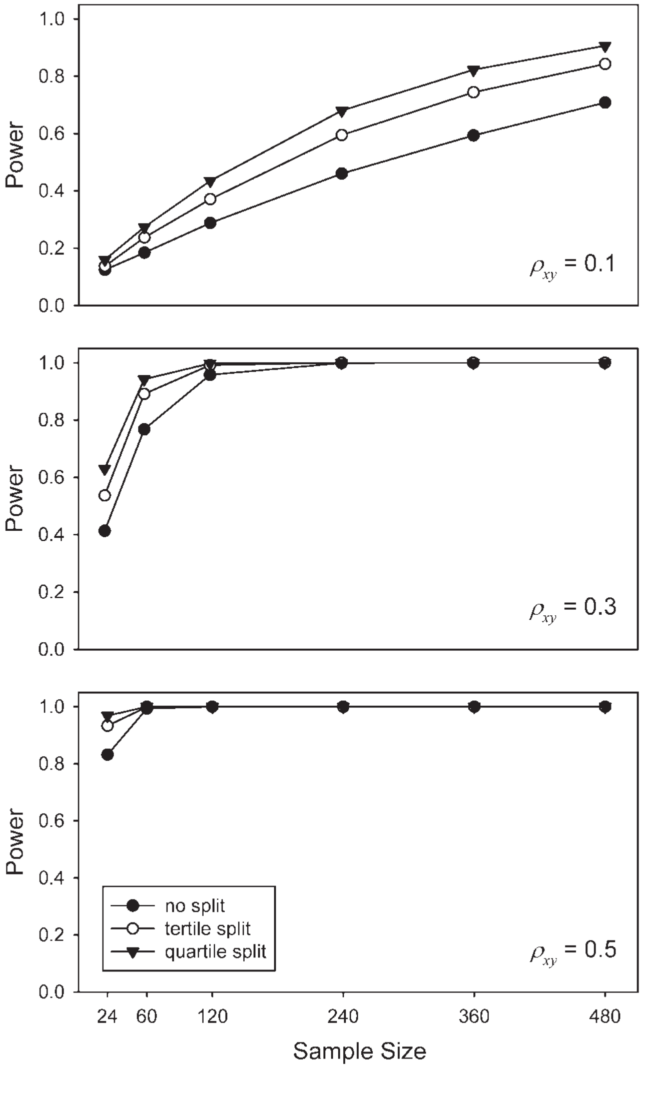

Figure 2. Empirical power of the test of $\rho_{xy} = 0$ when upper and lower halves, tertiles, and quartiles of x are used. Each point in a given panel represents the proportion of tests (out of 10,000) significant at α = .05. Sample size represents the number of cases collected from a larger sample of size $N(1 - p_o)^{-1}$, where p_o is the proportion of cases omitted on the basis of the extreme groups approach. Power increases with sample size, population correlation (ρ_{xy}), and for a given sample size (vertical slice), with the extremity of the split.

of a given size or to increase the power to detect an effect, EGA often may be implemented for no theoretical or empirical reason beyond the fact that it lowers p values to levels generally regarded as significant. We recognize that increasing the likelihood of statistical significance is, in one sense, the purpose of EGA; however, we strongly emphasize that modern recommendations regarding statistical methodology and reporting of results characterize practical significance or effect size as ultimately more important than achieving statistical significance (Wilkinson & the APA Task Force on Statistical Inference, 1999). Finally, it is important to emphasize that the power improvement associated with EGA does not apply when data are gathered from across the full range of x and then discarded from the middle of the x distribution (a procedure we describe as *post hoc subgrouping*; see the Recommendations for Practice section). The power benefit applies only when one compares EGA with the analysis of full-range data using the same sample size.

EGA and Effect Size

Effect size refers to the magnitude of an effect, for example the degree of linear dependence or amount of shared variance. Estimates of effect size can be classified as unstandardized or standardized. Unstandardized effect size estimates reflect the magnitude of an effect in raw units of whatever is being measured. For example, unstandardized regression weights reflect the degree of linear relationship in units of the dependent variable. Standardized effect size estimates (such as r_{xy}, R^2, η^2, ω^2, and Cohen's d) are expressed in common metrics unrelated to the raw scales of measurement of the observed variables. Thus, correlations express the degree of linear dependency between x and y on a scale ranging from -1.0 to $+1.0$ regardless of the metrics of x and y, and Cohen's d expresses mean differences in standard deviation units.

Measures of both unstandardized and standardized effect size are valuable tools in psychological research. However, even though standardized effect size measures are used widely in psychology, some authors have pointed out limitations. For example, Lenth (2001) pointed out that it is all too easy for researchers to fall into the trap of inappropriately using a particular value of a standardized effect size measure (e.g., Cohen's $d = 0.5$) as a target in a priori power analysis or sample size determination. P. Cohen, Cohen, Aiken, and West (1999) noted that standardized scores can be misleading when they are derived from samples of convenience that limit their generalizability or are based on skewed distributions. They further point out that commonly used benchmarks for standardized effect size measures may be misleading; for example, a Cohen's d of 0.2 may represent a small effect in many contexts but a large effect in others.

These are not inherent problems with the measures themselves but rather are potential misuses sometimes perpetrated by uninformed researchers. Standardized effect size estimates used responsibly can be important indicators of the extent or magnitude of an effect. They express degrees of relationship and differences in easily understood common metrics. When extreme groups are analyzed, standardized effect size tends to be "inflated dramatically" (Humphreys, 1985, p. 15), which in turn is associated with increased power. However, as Feldt (1961) pointed out, EGA should be used only to decide upon the presence of a linear effect in the population, not its strength (see also Pitts, 1993). Echoing the concerns of McNemar (1960), Feldt pointed out, "such a methodology is almost certain to be abused, for it can easily lead the experimenter to exaggerate the importance of trivial results" (p. 314).

To illustrate the effect EGA can have on standardized effect size, we conducted a simulation study. We generated data corresponding to three proportions of data omitted from the x distribution (.00, .33, .50). Using a constant sample size of 1,000, we simulated 51,000 bivariate data sets corresponding to values of ρ_{xy} ranging from 0.00 to 1.00 in steps of 0.02 (i.e., 1,000 data sets per value of ρ_{xy}). The mean sample correlation was computed for each value of ρ_{xy}, and the difference between sample correlations based on either tertile- or quartile-split data and full-range data were computed. These differences, plotted in Figure 3, represent the average gain in standardized effect size for values of ρ_{xy}, ranging from small to large for tertile and quartile splits, assuming bivariate normality and a linear relationship between x and y. The gain in standardized effect size can be quite large, especially for modest values of ρ_{xy}. Near the extremes of the range of ρ_{xy}, there is little change in standardized effect size after conducting EGA. For all values of ρ_{xy} between these extremes, however, the discrepancy between full-range data and data subjected to EGA will be somewhat further from zero. The slight asymmetry and irregularity in Figure 3 can be attributed to differences in the variability of correlation coefficients for different values of ρ_{xy} (as ρ_{xy} increases, the standard error of r_{xy} decreases).

Of course, improving effect size is one of the motivations for using EGA. However, it is inappropriate to interpret measures of standardized effect size as descriptive of population effects if they were derived from analyses performed on extreme groups. It is also inappropriate to statistically compare such estimates to each other or to apply meta-analysis to standardized effect size estimates based on extreme groups data. Because EGA almost always results in upwardly biased estimates of standardized effect size, to use EGA is to misrepresent the practical significance of effects in the population by inflating those estimates. In light of recent work on the appropriateness of unstandardized relative to standardized effect size measures (Bond, Wiitala, & Richard, 2003), we recommend researchers either use full-range data to estimate standardized effect size or report raw regression weights in lieu of r_{xy} when appropriate. We

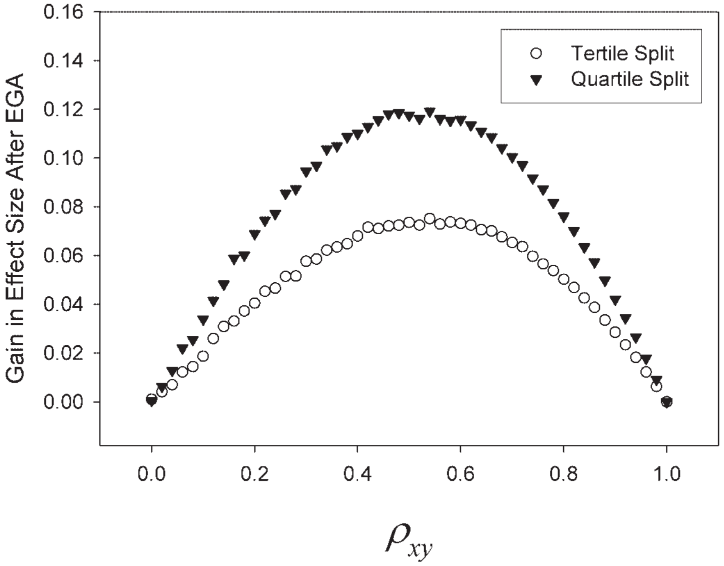

Figure 3. Mean difference between extreme-group and full-data sample correlations for tertile-split and quartile-split data. These curves represent the average gain in standardized effect size after tertile and quartile splits for values of population correlation (ρ_{xy}) ranging from small to large. The unevenness is attributable to sampling error in r_{xy}, which increases as r_{xy} decreases in magnitude. EGA = extreme groups approach.

recognize that measuring strength of association, and hence standardized effect size, is not always the scientist's goal. However, in many instances, such as when the degree of linear relationship or proportion of explained variance is the primary quantity of interest, selection of extreme groups should be avoided. The effect of EGA on effect size is particularly important in light of recent recommendations by Wilkinson and the APA Task Force on Statistical Inference (1999), which urged researchers to "always present effect sizes for primary outcomes" (p. 599). EGA thus has the potential to compromise some of the most important information reported in research when that information is communicated in the form of standardized effect size.

If EGA has been used, one may obtain an estimate of the full-range population correlation by using a curtailment of range formula provided by K. Pearson (1903, p. 23) and Wherry (1984). These formulas are used mainly to correct the underestimation of r_{xy} when a restricted range of x is observed, often as a result of selection of individuals above or below some threshold. In the present context, the same formula can be used to correct the overestimation of a correlation coefficient when only the extreme ranges of x are observed (Wherry, 1984, provides an example on p. 50). Although this is a potentially valid use of correction for range restriction, we have not seen such a correction used in any recent application of EGA.

Our observations on standardized effect size inflation may be pertinent to experimental designs more generally. Experiments with manipulated independent variables may produce misleading standardized effect sizes. For example, a researcher manipulating the variable *distraction* may use two extreme levels of distraction (e.g., no distraction versus high distraction). Interpretation of any standardized estimates of the magnitude of the effect in the population would necessarily be limited to the chosen levels of distraction and may or may not apply to other levels. This line of inquiry is interesting in its own right and deserving of future research.

EGA and Reliability

One rationale for EGA often mentioned in informal discussion with colleagues is that it improves power not only by enhancing standardized effect size but also by removing influences (such as unreliability in the middle of the distribution of x) that obscure an effect that really is present. Hence, selecting cases from the extremes of the distribution of x is thought to increase the reliability of a scale. In fact, as we show, EGA usually results in the omission of the most, not the least, reliable scores.

The most important concern here is whether it is generally true that measures are less reliable in the central part of the score distribution than in the extremes. It may be tempting to address the question of reliability by obtaining estimates of reliability before and after deleting data from the middle of the distribution. If reliability seems to increase after EGA, the researcher may want to conclude that using EGA was the proper thing to do. However, such a conclusion would be invalid; the apparent increase in reliability can be shown to be an artifact of the EGA procedure. Specifically, from the perspective of classical test theory, reliability is the

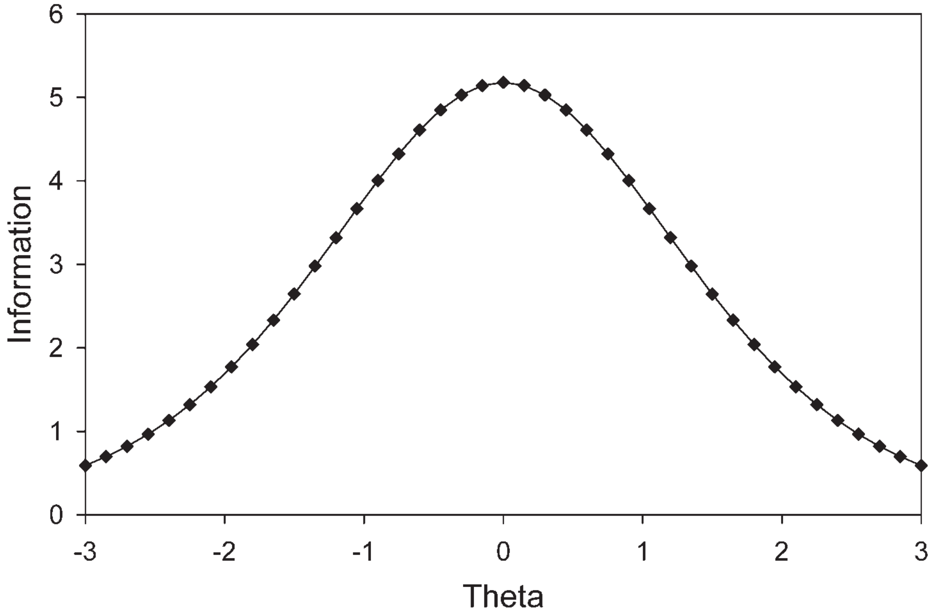

Figure 4. The item response theory information function for an arbitrary 15-item scale. Theta is the unobserved trait. Information is approximately the reciprocal of the error variance at each value of theta.

ratio of true variance to observed variance, where observed variance is the sum of true variance plus error variance. By deleting the middle of the distribution, EGA serves to increase true variance while leaving error variance unchanged, thus increasing reliability artifactually. This apparent increase in reliability is not due to elimination of less reliable scores, but rather to this manipulation of true variance. Therefore, simply computing reliability estimates before and after EGA is not informative regarding this question. True score theory (classical test theory) is not rich enough theoretically for answering questions about reliability in various regions of either the observed or true-score distributions. Item response theory (IRT), on the other hand, provides a more complete and appropriate framework for examining issues of relative reliability in different regions of a score distribution. Under the assumption of a normal distribution defining the latent trait being measured (θ), it is fairly easy to derive the reliability of observed scores in various regions of the latent trait distribution.

IRT defines a function relating the latent trait to item responses. One common function, the 2-parameter logistic model, has two item parameters: *a*, the item discrimination, and *b*, the item difficulty (or item endorsability). Some general guidelines for IRT are (a) reliable items (.3 or higher) have *a* values greater than or equal to 1 and (b) medium-difficulty, or medium-endorsable, items have *b* values in the vicinity of 0 (difficult items, or items that are infrequently endorsed, have *b* values greater than or equal to 1, and the *b* values for easy, or frequently endorsed, items are less than or equal to -1).

Researchers can quickly and visually determine where a measure is reliable and where it is not by using the information function, $I(\theta)$, of IRT. The information function for a collection of test items is the square of the standardized slope of the regression of the test score on the latent trait being measured.[3] Where this slope is steep, the true scores on the test are changing rapidly with changes in the latent trait, and hence the scores are reliable. In regions of the latent trait where the regression slope is shallow, true scores change very little with changes in theta; consequently, the test is not reliable for measuring this part of the latent trait distribution.

Consider a measure with 15 items; all *a* values are equal to unity in order to emulate a personality measure. The *b* values and their frequencies are –.50(2), –.25(2), 0(7), .25(2), and .50(2). The information function for this measure is given in Figure 4, and reliability information is reported in Table 1. It can be shown, by application of the delta theorem from statistics, that the information function $I(\theta)$ for a test composed of binary items can be interpreted as the local true-over-error variance for the observed score; therefore, a local reliability coefficient is $I(\theta)/[I(\theta) + 1]$.

[3] In more detail, the information function for the number-correct score at a fixed value of the latent trait (theta) is equal to $\frac{(\text{slope of score on } \theta)^2}{(\text{conditional variance of score on } \theta)}$. If the information function is used for specifying the precision of measurement (as indexed by the conditional standard error of measurement), no assumptions about the distribution of the latent trait are necessary.

Table 1
Reliability Information Derivable From IRT Information Function

	Portion of latent distribution			
Quantity	Lower	Middle	Upper	Total
Reliability	0.35	0.42	0.35	0.81
True variance	1.25	2.51	1.25	12.01
Error variance	2.28	3.49	2.28	2.88
Mean scale score	3.01	7.50	11.99	7.50

Note. IRT = item response theory.

The reciprocal of this same information function may also be interpreted as the conditional error variance of the maximum-likelihood estimate of θ. It is clear from this information function that observed scores are most reliable in the middle of the theta distribution and that observed scores will have low reliability in the extremes. It is very important to note that by examining the distribution of *b* values, one can determine where a measure will be most accurate: lower *b* values will produce higher reliability at the lower end of the latent distribution, and higher *b* values will result in higher reliability at the upper end of the latent distribution. If we assume that theta is normally distributed, the values in Table 1 can be obtained by numerically integrating the conditional distributions. The overall reliability is .81. The reliability in the middle half of the distribution ($-\ .67\sigma$ to $+\ .67\sigma$) equals 0.42. (This reduction in reliability is due principally to a decrease in true variance rather than to an increase in error variance.) In both the lower and upper 25% of the theta distribution, the observed scores have reliability equal to 0.35; in both cases the decreases in reliability are due mainly, not to increased error variance, but rather to diminished true variance. The reliability in the combined extremes is 0.88. This illustrative example makes it clear that for measures composed of moderately endorsable personality items (or cognitive items of moderate difficulty), there is likely to be greater reliability in the central portion of the distribution of the variable being measured (latent or observed). In order for the reliability of the extremes to increase as a result of EGA, the distribution of *b* values would need to follow an extremely unlikely U-shaped distribution, such that the pool of items consisted almost entirely of very low-endorsability or very high-endorsability items. In addition, the reliability in each of the extremes will never be equal to the reliability of a test in the entire population. Moreover, it is not true that EGA increases reliability only when the *b* values are clustered in the two extremes and if the extremes are combined into a single population; the true variance spuriously increases because of this concatenation.

One important aspect of the above IRT-based example is that the latent distribution rather than the observed distribution was partitioned. In practice, of course, EGA is applied to an observed distribution. However, it is quite problematic for researchers to investigate reliability when selecting scores on the observed distribution. Such selection induces nonzero correlations between true scores and errors, often highly negative. The existence of such correlations plays havoc with definitions of reliability and yields different estimates of reliability depending on which definition is used. Selecting on the latent distribution avoids this problem and should not yield misleading findings. As an illustration of the problems caused if one were to partition subpopulations through use of observed scores, in the upper 25% subpopulation, on the basis of observed scores in the example above, the reliability varies between −0.23 (for the $1\ -\ (\sigma_e^2/\sigma_x^2)$ definition of reliability) and 1.75 (for the $\sigma_{True}^2/\sigma_x^2$ definition). This discrepancy is because true and error scores are correlated –.62 in this subpopulation in the example. We believe the results of the demonstration above are sufficiently strong to make the case that the notion of middle scores being less reliable is not sensible in practice. Discussion of reliability at all in the context of EGA is problematic. For any selected observed score, true scores and errors are negatively correlated. For any selected latent score, the problem of correlated true scores and errors is avoided, but typically researchers do not have access to latent scores.

The issues considered here also relate to the question of the effect of EGA on statistical power. Reliability of the dependent variable is not directly related to the power of a statistical test (Nicewander & Price, 1983). The magnitude of within-group variance (the sum of true and error variances) is the major determinant of power. Control of individual differences is one of the goals of experimental design, and this concept translates directly into decreasing true variance. EGA is a form of experimental design, and in cases such as the one illustrated above, individual differences are reduced. The true variance in each extreme is much smaller than in the middle group; as a result, the average within-group variance is diminished. Thus, when EGA grants an advantage in terms of power, it is not likely a result of ridding the data of the unreliable middle, but rather of decreasing within-group variance by reducing true variance.

It was not our intent to prove anything in general about the reliability of measurement in various locales of a latent distribution. Our intent was to show, by application of a simple IRT model for binary items, that the middle of a distribution is not necessarily where scores are least reliable. Furthermore, our example made clear that low reliability is as apt to result from low true variance as it is from excessive error variance. The information function for binary items indicates where the observed scores on a measure will be most reliable. For more complicated IRT models—for example the partial credit model used for Likert and other items requiring multicategory responses—one may use the information function for the more complex model to indicate where the measure is most and least reliable. Again, for the case of more complex IRT models, the region(s) of the latent distribution where measurement precision is highest can be controlled by choosing items on the basis of their location (or difficulty) parameters. It is arguable that most measures are composed of items located in the central portion of the distribution of the latent trait being measured; therefore, many instruments will be most reliable for those in the middle.

To summarize, we wish to emphasize that two issues are important with regard to reliability in the middle of a distribution. First, scores are typically not less reliable in the middle of a score distribution; central scores are almost always more reliable than are those in the extremes. Second, when EGA is applied, reliability estimates for the combined extremes are typically elevated with respect to reliability for the entire distribution, but this is an artifact of the EGA procedure, attributable to an artifactual increase in true score variance. Application of IRT permits the appropriate assessment of reliability in different regions of a distribution and recognition of the effects of EGA on relevant variances.

EGA and Model Misspecification

The use of EGA involves the implicit assumption that the form of the x–y relationship across the range of x is the same as that in the extreme groups. The validity of the results rests in part on the correctness of this assumption; violations of this assumption are not difficult to envision and are probably quite common in practice (J. Cohen, Cohen, West, & Aiken, 2003). For example, a situation described by Sorrentino and Short (1977) involves a strong relationship between x and y at the tails of the x distribution but the absence of a stable relationship toward the middle. In other words, the relationship between x and y can sometimes be moderated nonlinearly by unmeasured situational variable(s) z. Of course, unstable relationships between x and y need not be limited to the middle of the x distribution. The true function relating y to x could be nonlinear in a variety of ways, but ignoring a segment of the population where nonlinearity may be evident could destroy the researcher's capability of finding it, essentially forcing linearity (McNemar, 1960). Thus, EGA has the potential to heighten the chances of model misspecification.

EGA would appear to increase the power of a subsequent test of the linear relationship between x and y but at the cost of rendering detection of any relationship other than linear improbable, in part because such effects would probably not be investigated by the researcher. This problem has been recognized before (e.g., Humphreys & Fleishman, 1974; Zedeck, 1971) but is rarely considered in practice. When the possibility of a nonlinear relationship between x and y cannot be legitimately dismissed, EGA should not be used. EGA should be considered "only when the assumption of linearity is strongly tenable" (Feldt, 1961, p. 314).

An alternative approach that reduces the odds of model misspecification is to avoid restricting attention to extreme groups data and instead fit different, possibly more complex models to full-range data (see, e.g., Bjerve & Doksum, 1993). Future research should weigh the practical simplicity of linear regression models applied to extreme groups data against the complexity of newly developed techniques applied to full-range data.

EGA and Group Assignment

Researchers may believe that a given construct is dichotomous by nature, despite the fact that the instrument used to assess it yields essentially continuous scores—in other words, it is often believed that latent taxons underlie observed individual differences in x. Arnold (1984), describing the frequent use of EGA in the context of moderator (interaction) analysis, characterized this rationale for EGA as "shift[ing] the measures into line with their underlying psychological values" (p. 222). For example, self-monitoring is an individual difference measure that taxometric analysis has suggested is composed of two distinct categories: low and high (Gangestad & Snyder, 1985; but see Miller & Thayer, 1989). Although taxometric procedures may be used to identify cutoff points, such techniques are rarely used for situations in which it is appropriate to do so. Instead, the existence of underlying taxons is often merely assumed. In this context, researchers sometimes use EGA to create two groups which, it is confidently (and conveniently) believed, represent these latent, categorical taxons. This use of EGA is commonly followed by dichotomization of the extreme scores into two categories. This may seem logical given early research by Feldt (1961). However, Alf and Abrahams (1975) showed that correlational analysis following the creation of extreme groups is more powerful than a t test comparing y means for extreme x groups. Their work shows that if EGA is to be used, researchers should not further reduce information by dichotomizing scores. Another reason not to dichotomize after EGA is that reliable information about individual differences is lost, as is the possibility to investigate nonlinear relationships between x and y. For additional problems associated with dichotomization of continuous scores see studies by J. Cohen (1983), Irwin and McClelland (2003), and MacCallum, Zhang, Preacher, and Rucker (2002).

Beyond the statistical problems associated with dichotomization, there are several problems with using EGA for the purpose of group assignment. First, it presumes that underlying classes or taxons exist—specifically, two taxons—and that observed scores would reflect taxon membership were it not for the presence of measurement error. We suggest that true dichotomies are not common in psychology and that even when they exist there are rigorous statistical procedures that can be used to determine whether the underlying distribution is discrete or continuous (e.g., Waller & Meehl, 1998).

Second, there is no guarantee that group assignment after dichotomization is accurate, yet there is no allowance for this possibility built into the statistical tests used after dichotomization. Thus, even when the presence of underlying classes can be justified, some individuals may still be misclassified. Misclassification represents a source of error above and beyond the usual measurement error assumed under classical test theory.

Third, use of EGA for purposes of group assignment involves the implicit assumption that the proportions of cases belonging to each group in the population are equal, yet it is doubtful that taxons are ever of equal size. For example, consider the case in which underlying taxons exist, but the population proportions for taxon membership are .25 (for low) and .75 (for high). A tertile split would include in the low group about 8.3% of those individuals who should have been classified as high. Even when there is a substantial middle group such that extreme groups in the sample consist only of those who really are low and high, unequal low and high population taxons cannot be well represented by equally sized extreme groups in the sample. Fourth, even if group assignment is accurate, there is no guarantee that group constituency will be stable across samples or across repeated measurements of the same sample.

To summarize, assignment of individual scores to arbitrary groups is problematic because it involves making possibly unwarranted assumptions about the existence of taxons, accuracy of group assignment, group size, and the stability of group membership.

EGA and Interaction Effects

Some researchers have suggested that creation of subgroups may be necessary in order to test some interaction or moderation hypotheses, particularly those involving differences in strength or degree of the relationship between variables *x* and *y* conditional on values of moderator *z* (Arnold, 1984; Sharma, Durand, & Gur-Arie, 1981). Subgroup analysis, as it is sometimes called in the marketing, medical, and industrial/organizational psychology literature, often proceeds by identifying a moderator variable *z*, creating groups on the basis of a median split or EGA, and comparing the relationship between variables *x* and *y* for the resulting *z* subgroups (Sharma et al., 1981). Investigation of the interaction effect proceeds by statistically comparing r_{xy} or R^2 from the two subgroups. However, comparison of quasi-arbitrary groups can be considered another form of model misspecification. Most interaction hypotheses are more appropriately tested by use of regression methods involving product terms as predictors (see Aiken & West, 1991). These methods are generally regarded as preferable because they (a) maintain the continuous nature of the potential moderator variable, (b) are often more appropriate for the hypotheses of interest (Stone & Hollenbeck, 1989; but see Arnold, 1982, 1984; Sharma et al., 1981), and (c) do not involve many of the costs associated with dichotomization.

Humphreys and Dachler (1969a) noted that selection of extreme groups on two variables followed by dichotomization and a two-way ANOVA is a problematic strategy. If the two variables subjected to EGA are correlated in the population, dichotomizing both variables results in a pseudo-orthogonal design, as the two independent variables will appear to be uncorrelated (orthogonal) in the sample even if they are actually correlated in the population. In such designs, effect sizes (standardized and unstandardized) and *p* values can be severely biased (Humphreys & Dachler, 1969a, 1969b; MacCallum et al., 2002).

EGA and Regression to the Mean

Campbell and Kenny (1999) noted that problems associated with the regression to the mean phenomenon are especially likely to occur with extreme groups. EGA is conducted on the presumption that extreme scores in the sample represent the extremes of the distribution of true scores in the population. However, those cases selected to be in the extremes in one instance may not be in the extremes if sampled at another time. The implication is that a phenomenon found to be statistically significant using EGA may be the result, at least in part, of a regression artifact.

To illustrate the problem of regression to the mean in the EGA context, we present a small demonstration. Two variables (x_1 and x_2) were generated from a bivariate normal distribution to represent repeated measures of selection variable *x* with $N = 1{,}000$. The test–retest correlation, a frequently used index of stability over time (test–retest reliability), was defined to be relatively high at $r_{xx} = .90$ in both the population and the sample. Separate tertile and quartile splits were performed on x_1 and x_2. For each type of split, cases were categorized into one of nine groups, depending on their pattern of extremity across two trials. The results reported in the upper half of Table 2 demonstrate that, for the tertile split condition, only about 81% of cases that were extreme at Time 1 maintained their extreme status at Time 2, whereas about 37% of the cases omitted at Time 1 were retained as extreme scorers at Time 2. For the quartile split condition, approximately 76% of extreme cases at Time 1 were also extreme at Time 2, whereas 24% of middle scorers at Time 1 became extreme at Time 2. Overall, little more than 75% of the cases maintained their extremity status (whether midrange or extreme) from Time 1 to Time 2. This demonstration was repeated with $r_{xx} = .80$, a more realistic value in many realms of psychological research. Regression to the mean is more pronounced when r_{xx} is smaller (see the lower half of Table 2). Overall, little more than two thirds of the cases maintained their extremity status from Time 1 to Time 2. If EGA were effective at retaining extreme scorers from Time 1 to Time 2, Table 2 would contain near-diagonal matrices. Thus, we

Table 2
Pattern of Extremity Across Repeated Trials

	Group at Time 2 (tertile split)			Group at Time 2 (quartile split)		
Group at Time 1	Low	Middle	High	Low	Middle	High
$r_{xx} = .90$						
Low	272	60	1	189	61	0
Middle	57	212	65	61	382	57
High	4	62	267	0	57	193
$r_{xx} = .80$						
Low	241	76	16	172	77	1
Middle	83	183	68	78	346	76
High	9	75	249	0	77	173

Note. The number in each cell represents the frequency of cases (out of 1,000) in a particular group at Time 1 that were in a particular group at Time 2.

should be careful about assuming that extreme scorers at Time 1 can be relied upon to remain extreme at Time 2, at least in examples such as those simulated here.

Potential Applications of EGA

Given the problems outlined in this article, we find the usefulness and appropriateness of EGA as a methodological technique to be somewhat limited. If EGA is used, greater power may result, but the researcher will also be obligated to (a) avoid interpreting inflated standardized effect size measures, (b) risk model misspecification by making possibly unwarranted assumptions about linearity and group membership, and (c) recognize that reliability is most likely reduced rather than increased. However, in this section we consider possible circumstances in which EGA may represent exemplary practice or the most appropriate course of action given constraints on study design, data collection, or the state of knowledge in a field. We hasten to add that typical uses of EGA are unlikely to have been the result of seriously weighing the advantages and disadvantages of the technique and that all of these potential uses bear close examination to determine whether they are in fact legitimate or useful in a given setting. Therefore, the following speculative suggestions should be considered directions for future research rather than endorsements for appropriate uses of EGA.

Cost-Efficiency and the Power to Detect Effects

The use of EGA may be a matter of necessity in situations when a researcher has limited resources and wishes to maximize the power for detecting the presence of an effect. Cost, in terms of either time or money, may not permit examination of the full range of data. In such situations, and with proper considerations, EGA may be a useful tool to improve the odds of detecting an effect, if it truly exists. In pilot studies and exploratory research in which little prior knowledge exists, EGA can be useful to detect general trends in the data. The exact functional form linking y to x may be immaterial in the early stages of research, as long as there is evidence of a relationship. However, we caution that in such situations the researcher must remain vigilant with regard to several facts.

First, using EGA in pilot studies will result in inflated standardized effect size estimates, which in turn have the potential to lead to false expectations when it comes time to conduct the study proper. The use of EGA to enhance the likelihood of detecting an effect also carries a real risk of model misspecification. The researcher must be able to support the assumption that the same model applicable to the extremes also applies to the omitted middle, an assumption that may be supportable on theoretical grounds or on the basis of previous research. In general, therefore, EGA may represent a valuable tool for determining (a) whether an effect exists and (b) the direction of an effect but not for determining the size of an effect (Brunswik, 1955; Cortina & DeShon, 1998; Feldt, 1961; Pitts, 1993), at least in standardized units. Only those effects reported in unstandardized units will remain unbiased after EGA is used.

Nonnormal Data

In many cases, data are nonnormally distributed or so severely skewed that standard parametric statistics may not be appropriate. For example, smoking behavior is commonly assessed by measuring the frequency of smoking over some period of time. Depending on the population, the smoking frequency distribution may be highly skewed. In such situations EGA could be used to create groups (e.g., nonsmoker and heavy smoker). Cureton (1957) and Fowler (1992) investigated the effects of violating the normality assumption on the power of *t* tests performed after EGA followed by dichotomization, finding that power is maximized when 27% (or slightly more) of the cases are retained in each tail. However, the consequences of using EGA under severe skewness or kurtosis, while maintaining the continuous nature of the data, has yet to be formally investigated. Procedures such as Poisson regression (J. Cohen et al., 2003), developed specifically for analysis of count data, would be more appropriate in many instances.

EGA and Interaction Effects

Humphreys and Dachler (1969a, 1969b) found that creating unrepresentative subgroups for use in ANOVA designs is problematic. On the other hand, in their investigation of difficulties associated with detecting interaction (moderation) effects in observational research, McClelland and Judd (1993) found that the optimal design for detecting interactions—that is, the situation in which statistical power for the test of the multiplicative effect of two variables predicting a third variable is maximized—is one in which there are jointly extreme scores in the distributions of predictor variables x and z. If sample size is fixed and a researcher is given the choice of where in the ranges of x and z to allocate subjects, the findings of McClelland and Judd (1993) imply that the extremes are the natural choice, because the result would be an optimal design and thus would possess more power to detect the interaction effect of interest relative to a design including data from the middle of the distributions of x or z. Indeed, McClelland and Judd (1993) and Pitts (1993) suggested that one way researchers can achieve a near-optimal design for detecting interactions is to oversample extreme observations. Oversampling represents a less extreme form of EGA in that midrange scores are still collected and retained, but extreme scores occur in numbers disproportionate to their natural frequencies. McClelland and Judd noted that unlike correlation coefficients, which are standardized indices of linear association, the expected values of unstandardized regression coefficients will not change after oversampling strategies like EGA (if linearity is assumed; see also Pitts, 1993). However, they cautioned that oversampling seriously inflates R^2. Thus, oversampling (or EGA) can be used to increase the power to detect interaction effects only if the assumption of linearity can be made, but no conclusions should be drawn about the standardized effect size in terms of the percentage of variance explained.

Recommendations for Practice

There are some productive, justifiable uses for EGA. EGA can be beneficial in terms of improving cost-efficiency. If sample size is limited, EGA often can be used to increase the power to detect an effect, if an effect exists to be found. Thus, EGA appears to be well-suited for use in pilot studies or in the exploratory phase of research, when the exact functional form of a relationship is unknown but there is reason to make conjectures about the existence and direction of an effect. EGA may be useful for the detection of hypothesized lower order interaction effects. In addition, unstandardized effect size estimates (such as raw regression weights) will be unbiased whether or not EGA is used. When the sample size is large enough to support the collection of data representative of the full range of x, the data permit the use of more sophisticated techniques, such as multilevel modeling, linear and nonlinear multiple regression, and structural equation modeling. Mixture modeling or taxometric analysis might be used in conjunction with these modeling techniques to investigate the presence or absence of distinct groups. When more information is available, richer knowledge can be extracted from data.

We recommend that researchers should, as the default choice, use traditional full-data correlation and regression approaches to analyze naturally continuous data when conditions allow it. As long as the assumptions of linearity, homoscedasticity, and residual normality are met (or at least reasonably approximated), parameter estimates obtained from ordinary regression and correlation approaches are unbiased and informative. Furthermore, inferences about the population can be drawn, predictions can be made, and effects can be compared. We believe the use of EGA should be rare rather than common. However, if a researcher finds it necessary or advantageous to use EGA, we encourage adhering to the following recommendations.

Justify the Use of EGA

Researchers should provide explicit justifications for the decision to use EGA. Was EGA used to improve cost efficiency? Was EGA used to increase power for the detection of some hypothesized effect? Before using EGA to increase power, researchers should demonstrate that problems of low power exist (e.g., because of a limited sample size, a small effect). Furthermore, can the assumption of a linear relationship between x and y safely be assumed? We strongly discourage the use of EGA as a mechanical tool that is used simply because it was used in the past or because it results in smaller *p* values. However, our reading of the literature indicates that (a) EGA is often used even when conditions permit collection of data across the full range of x, and (b) conclusions drawn from analyses after EGA is used can be easily overstated.

Do Not Dichotomize Data After Extreme Group Selection

Perhaps the single most important change that should be made to current use of EGA is to avoid dichotomizing data as part of the procedure. Alf and Abrahams (1975) demonstrated that after extreme groups have been selected, fitting a line to continuous data was more powerful than was a *t* test performed on grouped data. If x and y are related in a nonlinear fashion, dichotomization removes the possibility of identifying the true relationship (Humphreys & Fleishman, 1974). Furthermore, because power is actually lost following dichotomization, dichotomization is rarely a good idea (MacCallum et al., 2002).

Avoid Post Hoc Subgrouping

A different procedure, which we refer to as *post hoc subgrouping*, consists of obtaining data for both variables x and y for all individuals in a sample but analyzing data on y only for those individuals scoring at the extremes on variable x. This approach is used for reasons other than cost-efficiency, such as the unfounded belief that an analysis of the extremes improves statistical power by reducing error variance (thereby increasing reliability) or increasing effect size.

Unlike EGA, subgrouping does not improve cost-efficiency, because data on y have already been collected for all individuals. In addition, relative to analysis of the full sample, subgrouping usually lowers the power of subsequent hypothesis tests. In fact, if error variability can be assumed constant across the range of x, Alf and Abrahams (1975) showed analytically that power for the test of a null hypothesis that $\rho_{xy} = 0$ cannot be enhanced by removing cases. Post hoc subgrouping is similarly ill-advised when one attempts to detect interactions. McClelland and Judd (1993) demonstrated that adding a midrange value to an optimal design can only increase the power of the test of an interaction effect. It follows that removing such a value when present would decrease power, and they rightly characterized this procedure as "unwise" (p. 386) Given that there appear to be no clear advantages to subgrouping, we strongly recommend against the common practice of applying tertile and quartile splits to otherwise full data sets.

Conclusion

McNemar (1960), in his presidential address to the Western Psychological Association, noted that, "By the extreme groups method, everybody is kept in a state of blissful ignorance" (p. 298). By this he meant that if some data are ignored and the remaining data are assigned to groups via dichotomization, the researcher has no idea what the missing data could tell us, and that trivial results can easily be magnified out of proportion. We agree with McNemar's concerns, and we further caution that the conclusions that can be based on the results of EGA, although sometimes useful, are limited relative to those based on analysis of full-range, continuous data. Furthermore, the use of EGA encourages a small-scale, bivariate approach to solving scientific problems in situations in which multivariate analysis may be more appropriate. In addition, EGA almost always limits the usefulness of tests of nonlinear effects; the form of the relationship between x and y must remain in doubt if EGA is used.

On the other hand, it sometimes may be appropriate to make claims about the presence and general direction of a relationship even if its size and shape are debatable. If resources are limited or if research is still in the exploratory stage in which there is little prior research to guide theory development, EGA can sometimes be used to enhance the detectability of effects. If a researcher does not have reason to use EGA beyond the fact that it increases the odds of achieving statistical significance, we strongly caution against the use of EGA. Given the risks associated with EGA, we suggest that any implementation of its use should be accompanied by careful consideration and clear justifications. Furthermore, even if EGA appears justified, we urge researchers to incorporate the recommendations for using EGA suggested in this article, such as avoiding making claims beyond those supported by the data and keeping the data in their original, continuous form. We urge reviewers, editors, and consumers to consider the appropriateness of instances of EGA encountered in the literature.

References

Abrahams, N. M., & Alf, E. F., Jr. (1978). Relative costs and statistical power in the extreme groups approach. *Psychometrika, 43,* 11–17.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks, CA: Sage.

Alf, E. F., Jr., & Abrahams, N. M. (1975). The use of extreme groups in assessing relationships. *Psychometrika, 40,* 563–572.

Arnold, H. J. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance, 29,* 143–174.

Arnold, H. J. (1984). Testing moderator variable hypotheses: A reply to Stone and Hollenbeck. *Organizational Behavior and Human Performance, 34,* 214–224.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics, 5,* 207–212.

Bernichon, T., Cook, K. E., & Brown, J. D. (2003). Seeking self-evaluative feedback: The interactive role of global self-esteem and specific self-views. *Journal of Personality and Social Psychology, 84,* 194–204.

Binet, A. (1900). Attention et adaptation [Attention and adaptation]. *L'Année Psychologique, 6,* 248–404.

Bjerve, S., & Doksum, K. A. (1993). Correlation curves: Measures of association as functions of covariate values. *Annals of Statistics, 21,* 890–902.

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8,* 406–418.

Borich, G. D., & Godbout, R. C. (1974). Extreme groups designs and the calculation of statistical power. *Educational and Psychological Measurement, 34,* 663–675.

Brunswik, E. (1955). Representative design and probabilistic theory. *Psychological Review, 62,* 193–217.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34,* 315–346.

Cortina, J. M., & DeShon, R. P. (1998). Determining relative importance of predictors with the observational design. *Journal of Applied Psychology, 83,* 798–804.

Cross, S. E., Morris, M. L., & Gore, J. S. (2002). Thinking about oneself and others: The relational-interdependent self-construal and social cognition. *Journal of Personality and Social Psychology, 82,* 399–418.

Cureton, E. E. (1957). The upper and lower twenty-seven per cent rule. *Psychometrika, 22,* 293–296.

D'Agostino, R. B., & Cureton, E. E. (1975). The 27 percent rule revisited. *Educational and Psychological Measurement, 35,* 47–50.

Darvasi, A., & Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics, 85,* 353–359.

Deffenbacher, J. L., Huff, M. E., Lynch, R. S., Oetting, E. R., & Salvatore, N. F. (2000). Characteristics and treatment of high-anger drivers. *Journal of Counseling Psychology, 47,* 5–17.

Feldt, L. S. (1961). The use of extreme groups to test for the presence of a relationship. *Psychometrika, 26,* 307–316.

Flanagan, J. C. (1939). General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. *Journal of Educational Psychology, 30,* 674–680.

Fowler, R. L. (1992). Using the extreme groups strategy when measures are not normally distributed. *Applied Psychological Measurement, 16,* 249–259.

Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review, 92,* 317–349.

Garg, R. (1983). An empirical comparison of three strategies used in extreme group designs. *Educational and Psychological Measurement, 43,* 359–371.

Gibson, W. M., & Jowett, G. H. (1957). 'Three-group' regression analysis: Pt. I. Simple regression analysis. *Applied Statistics, 6,* 114–122.

Henshall, J. M., & Goddard, M. E. (1999). Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics, 151,* 885–894.

Humphreys, L. G. (1978). Research on individual differences requires correlational analysis, not ANOVA. *Intelligence, 2,* 1–5.

Humphreys, L. G. (1985). Correlations in psychological research. In D. K. Detterman (Ed.), *Current topics in human intelligence: Vol. 1. Research methodology* (pp. 3–24). Norwood, NJ: Ablex Publishing.

Humphreys, L. G., & Dachler, H. P. (1969a). Jensen's theory of intelligence. *Journal of Educational Psychology, 60,* 419–426.

Humphreys, L. G., & Dachler, H. P. (1969b). Jensen's theory of intelligence: A rebuttal. *Journal of Educational Psychology, 60,* 432–433.

Humphreys, L. G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology, 66,* 464–472.

Iacobucci, D. (Ed.). (2001). Methodological and statistical concerns of the experimental behavioral researcher [Special issue]. *Journal of Consumer Psychology, 10*(1–2).

Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research, 40,* 366–371.

Jensen, M. B. (1928). Objective differentiation between three groups in education (teachers, research workers, and administrators). *Genetic Psychology Monographs: Child Behavior, Differential and Genetic Psychology, 3,* 333–454.

Kagan, J., Snidman, N., & Arcus, D. (1998). The value of extreme groups. In R. B. Cairns, L. R. Bergman, & J. Kagan (Eds.), *Methods and models for studying the individual: Essays in honor of Marian Radke-Yarrow* (pp. 65–82). Thousand Oaks, CA: Sage.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 179–182.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17–24.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61,* 213–218.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55,* 187–193.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40.

McCabe, G. P. (1980). Use of the 27% rule in experimental design. *Communications in Statistics, A9,* 765–776.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390.

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist, 15,* 295–300.

Miller, M. L., & Thayer, J. F. (1989). On the existence of discrete classes in personality: Is self-monitoring the correct joint to carve? *Journal of Personality and Social Psychology, 57,* 143–155.

Mosteller, F. (1946). On some useful "inefficient" statistics. *Annals of Mathematical Statistics, 17,* 377–408.

Muranty, H., & Goffinet, B. (1997). Selective genotyping for location and estimation of the effect of a quantitative trait locus. *Biometrics, 53,* 629–643.

Nair, K. R., & Banerjee, K. S. (1942). A note on fitting of straight lines if both variables are subject to error. *Sankhyā, 6,* 331.

Nair, K. R., & Shrivastava, M. P. (1942). On a simple method of curve fitting. *Sankhyā, 6,* 121–132.

Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. *Psychological Bulletin, 94,* 524–533.

Pearson, E. S., & Hartley, H. O. (1956). *Biometrika tables for statisticians* (Vol. 1). Cambridge, England: Cambridge University Press.

Pearson, K. (1903). Mathematical contributions to the theory of evolution: XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London: Series A, 200,* 1–66.

Peters, C. C. (1941). A technique for correlating measurable traits with freely observed social behaviors. *Psychometrika, 6,* 209–219.

Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases.* New York: McGraw-Hill.

Pitts, S. C. (1993). *The utility of extreme groups analysis to detect interactions among correlated predictor variables.* Unpublished master's thesis, Arizona State University, Tempe.

Pontari, B. A., & Schlenker, B. R. (2000). The influence of cognitive load on self-presentation: Can cognitive busyness help as well as harm social performance? *Journal of Personality and Social Psychology, 78,* 1092–1108.

Ross, J., & Lumsden, J. (1964). Comment on Feldt's "use of extreme groups." *Psychometrika, 29,* 207–209.

Ross, J., & Weitzman, R. A. (1964). The twenty-seven per cent rule. *Annals of Mathematical Statistics, 35,* 214–221.

Sharma, S., Durand, R. M., & Gur-Arie, O. (1981). Identification and analysis of moderator variables. *Journal of Marketing Research, 18,* 291–300.

Sorrentino, R. M., & Short, J.-A. C. (1977). The case of the mysterious moderates: Why motives sometimes fail to predict behavior. *Journal of Personality and Social Psychology, 35,* 478–484.

Stone, E. F., & Hollenbeck, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: Empirical evidence and related matters. *Journal of Applied Psychology, 74,* 3–10.

Torgesen, J. K. (1991). Subtypes as prototypes: Extended studies of rationally defined extreme groups. In L. Vernon-Feagans & E. J. Short (Eds.), *Subtypes of learning disabilities: Theoretical perspectives and research* (pp. 229–246). Hillsdale, NJ: Erlbaum.

Verplanken, B., & Holland, R. W. (2002). Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology, 82,* 434–447.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics, 11,* 284–300.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* Thousand Oaks, CA: Sage.

Wherry, R. J. (1984). *Contributions to correlational analysis.* New York: Academic Press.

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychological Bulletin, 76,* 295–310.

Received November 10, 2003
Revision received December 13, 2004
Accepted December 14, 2004 ■