When Effect Sizes Disagree: The Case of *r* and *d*

Robert E. McGrath
Fairleigh Dickinson University

Gregory J. Meyer
University of Toledo

The increased use of effect sizes in single studies and meta-analyses raises new questions about statistical inference. Choice of an effect-size index can have a substantial impact on the interpretation of findings. The authors demonstrate the issue by focusing on two popular effect-size measures, the correlation coefficient and the standardized mean difference (e.g., Cohen's *d* or Hedges's *g*), both of which can be used when one variable is dichotomous and the other is quantitative. Although the indices are often practically interchangeable, differences in sensitivity to the base rate or variance of the dichotomous variable can alter conclusions about the magnitude of an effect depending on which statistic is used. Because neither statistic is universally superior, researchers should explicitly consider the importance of base rates to formulate correct inferences and justify the selection of a primary effect-size statistic.

Keywords: effect-size estimation, base rate, correlation coefficient

In recent years, behavioral researchers have witnessed an important change in what is considered optimal statistical practice. With growing awareness of the differences between statistical and practical significance, the importance of power analysis for significance testing, meta-analysis as an integrative strategy, and the limitations of significance testing (e.g., Harlow, Mulaik, & Steiger, 1997; Thompson, 2002), recommendations for incorporating effect-size estimates into statistical analyses have become more definitive. For example, the fourth edition of the *Publication Manual of the American Psychological Association* (American Psychological Association [APA], 1994) "encouraged" authors to report effect sizes in statistical analyses (p. 18). By 1999, Leland Wilkinson and the APA Task Force on Statistical Inference wrote "always present effect sizes for primary outcomes," and "we must stress . . . that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (p. 599). In response to this recommendation, the most recent edition of the APA *Publication Manual* (APA, 2001, p. 25) indicates the reporting of effect sizes is "almost always necessary." More than 20 journals in the field of behavioral research now require authors to report effect-size statistics, at least for key statistical analyses (a list is provided at http://www.

Computational Issues

Cohen's *d*

r_{pb} conceptualizes relationships in terms of the degree to which variability in the quantitative variable and the dichotomous variable overlap.

One standard formula for the point-biserial correlation as a descriptive rather than inferential statistic is as follows:

$$r_{pb} = \frac{(\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2})}{S_Y}\sqrt{p_1 p_2}. \qquad (5)$$

S_Y is the standard deviation generated by dividing the total sums of squares for the quantitative variable by N. When $\bar{Y}_{\cdot 1} \neq \bar{Y}_{\cdot 2}$, S_Y is larger than S_{pooled}, the standard deviation used to compute d (Equation 2), and the size of the difference between the two standard deviations is directly related to the size of the difference between the means (demonstrated in the Appendix). As a result, the correlation is bounded within the interval -1.00 to 1.00.[3] The formula also includes the terms p_1 and p_2, which indicate the base rates or proportions of participants in each of the dichotomous variable groups, with $p_2 = 1 - p_1$.

The Effect of Base-Rate Inequalities

The reason d and r_{pb} can lead to different conclusions can be demonstrated several different ways. As the difference between p_1 and p_2 in Equation 5 increases, their product becomes smaller, so r_{pb} decreases. Because p_1 and p_2 are not part of the formula for d, the latter statistic is unaffected by base-rate disparities. As a result, d and r_{pb} differ markedly in terms of the degree to which they are affected by the base rate for the two values of the dichotomous variable. Thus, r_{pb} can be understood as a base-rate-sensitive effect-size measure, whereas d is base-rate-insensitive.

This difference in sensitivity to base rates can also be stated in terms of the variance of the dichotomous variable. Because the variance of this variable is a function of the product of the base rates (i.e., with the dichotomous groups coded as two consecutive numbers such as 0 and 1, $S_X^2 = p_1 p_2$ and $S_X = \sqrt{p_1 p_2}$), variance is maximized when $p_1 = p_2 = .50$. As the proportions become more discrepant, the variance of the dichotomous variable becomes smaller (see Figure 1, left vertical axis), resulting in a decline in the value of the correlation similar to that resulting from range restriction. Thus, rather than saying r_{pb} is base-rate-sensitive and d is base-rate-insensitive, one could just as readily state that r_{pb} is a variance-sensitive effect-size measure, whereas d is variance-insensitive. In this case, it is important to remember that the variance referred to is that of the dichotomous variable not the within-group or total variance for the quantitative variable. Goodman (1991) also suggested the terms marginal-dependent and marginal-free to represent the two classes of statistics.

In pursuit of making the difference between d and r_{pb} even clearer, the standard formulas can be modified to illustrate the two critical distinctions:

$$d = \frac{(\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2})}{\sqrt{S^2_{pooled}}} \qquad (6)$$

and

[3] The true possible range of the point-biserial correlation is actually smaller. No correlation can reach a value of 1.00 unless the two variables have the same distribution. Because quantitative and dichotomous variables by definition have different distributions, the true range for the point-biserial correlation is always less than 1.00 to -1.00 and varies depending on the distribution of the quantitative variable. For example, Nunnally and Bernstein (1994) reported the point-biserial correlation is restricted to the interval from $-$

$$r_{pb} = (\bar{Y}_{\cdot 1} - \bar{}$$

a larger r_{pb} value. For example, when $p_1 = .75$, a d of 0.80 is associated with r_{pb} of .33, whereas a d of 0.50 is associated with r_{pb} of .21. Thus, when p_1 is constant, the rank ordering of effect sizes is preserved across the two measures.

This relationship no longer exists when p_1 varies across analyses. For instance, consider the first three rows in Table 2 when $p_1 = .98$, $d = 0.80$ is associated with $r_{pb} = .11$. However, as p_1 approaches .50, r_{pb} increases so that $r_{pb} = .33$ when $p_1 = .75$, and $r_{pb} = .37$ when $p_1 = .50$. Even though d did not change, r_{pb} increased when the difference in base rates was less extreme. Furthermore, what is often considered a large d value (i.e., 0.80; Cohen, 1988) is associated with a small value for r (i.e., .11), when the probability of one of the two dichotomous values is only .02. A base rate of .02 (2 cases per 100) may seem like an extremely rare outcome, but in fact it is not. For instance, many psychiatric conditions have a prevalence of .02 or less in the general population, including dysthymia, agoraphobia, panic disorder, bipolar disorder, schizophrenia, any drug use disorder, or any specific personality disorder (Narrow, Rae, Robins, & Regier, 2002; Torgersen, Kringlen, & Cramer, 2001). The same is true for numerous medical conditions. For example, a recent study found that 2.3% of older males with normal levels of prostate-specific antigen had a serious form of prostate cancer upon biopsy (Thompson et al., 2004). It is also likely that many social and experimental phenomena commonly studied by psychologists are similarly infrequent, though—as we discuss below—it is often difficult to estimate the true frequency of these events.

Table 2 demonstrates another impediment to achieving comparable results with r_{pb} and d, though not for mathematical reasons. Many users of Cohen's (1988) benchmarks seem unaware that those for the correlation coefficient and d are not strictly equivalent, because Cohen's generally cited benchmarks for the correlation were intended for the infrequently used biserial correlation rather than for the point biserial. This creates a slight advantage for d over r in terms of the characterization of effect sizes when those benchmarks are used for other types of correlation coefficients. For example, as demonstrated in the table, when base rates are equal, the d value Cohen suggested as large (0.80) corresponds to an r_{pb} value of .37, far less than his commonly cited benchmark for a large r value (.50). To achieve comparability between r_{pb} and d when base rates are equal, the benchmarks for small, medium, and large correlations would need to be changed to .10, .24, and .37, respectively (Cohen, 1988, pp. 22, 82; Lipsey & Wilson, 2001, p. 147). Alternatively, to equate d with

by only about 10%. However, it is important to recognize that for purposes of enhancing power, researchers often oversample target cases or use equal-sized target and control groups, which may seriously underestimate the degree of base-rate inequality in the population. For instance, it is questionable whether 37% represents a reasonable estimate of how frequently cognitive impairment occurs in many applied settings. If testing was being conducted in an educational setting in which just 10% of the children were expected to have some form of cognitive impairment, the validity coefficient should drop from $r = .32$ to $r = .23$.

A similar analysis can be applied to the experimental study of psychological phenomena, though the comparison is often complicated by the lack of information about the true base rates for the events studied. To illustrate, we provide an example from social psychology. Carlson, Marcus-Newhall, and Miller (1990) presented a meta-analysis of studies investigating whether aggressive cues facilitate aggressive responding in negatively toned situations. They found that aggressive responding was greater when a weapon was present than when it was not, as long as there was no evidence the participants were aware of the research hypothesis; the mean *d* value was 0.31. Most of the studies they cited used equally sized groups, even those Anderson, Lindsay, and Bushman (1999) later classified as field studies that should generalize to everyday life. Because these experiments look poor (Brennan & Prediger, 1981; Spitznagel & Helzer, 1985; Zwick, 1988). Others have contended in response that base-rate sensitivity is appropriate to a reliability statistic (Bartko, 1991; Shrout, Spitzer, & Fleiss, 1987). As the base-rate inequality increases, total variance decreases. This means that measurement error variance will tend to increase as a proportion of the total variance, and reliability in fact decreases.

Haddock, Rindskopf, and Shadish (1998) argued for the odds ratio on the basis of its insensitivity to the distribution of dichotomous variables. Some researchers have conducted comparisons of effect-size measures under the assumption that effect sizes should not be affected by variable distributions (Hunter, 1973; von Eye & Mun, 2003); others have not made this assumption (Costner, 1965; Kraemer et al., 1999). It is not surprising that the former have criticized the correlation coefficient, whereas the latter have been more supportive of its use. For example, Kraemer et al. (1999) were troubled by the odds ratio's indirect relationship to power, a criticism that as noted above can be leveled at *d* and at all other base-rate-insensitive effect-size measures.[6] Complicating matters is the possibility, to be discussed below, that the purposes of the analysis may be an important factor in deciding between base-rate-sensitive and insensitive statistics.

is a better indicator of the *p* value resulting from the corresponding significance test.

It also suggested to us that standard discussions of the relationship between *d* and power have been remiss in overlooking the issue of base rates. Most textbooks acknowledge the impact of total sample size and alpha level on power. We know of none that includes extreme base rates in a dichotomous variable in the list of moderators of the relationship between effect size and power for base-rate-insensitive statistics, even though it often might be more important in practice than alpha level. Rosnow, Rosenthal, & Rubin (2000; Equation 10) provide an index of the relationship between base-rate inequality and subsequent loss of power, loss $= 1 - (n_h/\bar{n})$, where n_h is the harmonic mean of the group sizes and $\bar{n}$ is their arithmetic mean. When converted to base-rate notation, the formula indicates that the relative loss of statistical power from unequal sample sizes is $1 - 4p_1 \, p_2$. Thus, when base rates are equal, there is no loss of power $[1 - 4(.5)\ (.5) = 0]$. However, when 95% of the participants are in one group and 5% are in the other, power declines by 81% $[1 - 4(.95)\ (.05) = .81]$. Because variance in the dichotomous variable is determined by $p_1 \, p_2$, it also can be seen that, all other factors being equal, power is directly proportional to the variances indicated in Figure 1. That is, power is at a maximum (and relative loss is at a minimum) in the center of the figure when p_1 and $p_2 = .50$ but drops precipitously as the base rates diverge. Thus, Figure 1 simultaneously illustrates constraints on the size of r_{pb} (left vertical axis) and the relative power associated with *d* (right vertical axis). Discussions of power should straightforwardly indicate the role of base rates in power analysis. In the absence of this information, the uninformed user can easily overestimate the power of the study based on *d*.

2. *r* is a more flexible statistic.

The correlation coefficient can be computed for any combination of dichotomous and quantitative variables. This is an extremely useful characteristic when attempting to make comparisons across a variety of study designs, as is sometimes the case in meta-analysis. Rosenthal (1991) and colleagues (Rosenthal et al., 2000) have provided methods for the use of *r* as a general indicator of magnitude that is applicable in almost any study. This sort of flexibility is unusual among statistics. Though *d* can be used when both variables are dichotomous (Haddock et al., 1998), it cannot be used when both are quantitative.

The greater practical flexibility of *r* corresponds to the broader relevance of the concept of association or relationship versus group differences. In almost any circumstance in which a researcher is interested in considering two variables in conjunction with one another, that interest can be conceptualized in terms of an association between the variables. In contrast, the concept of differences between groups represents a special case of understanding contingent relationships.

3. *r* is integral to general linear models.

r is a cornerstone of multiple regression statistics, including the standard error of estimate, the regression coefficient, and the index of association. Indeed, *r* is central to the general linear model in all its forms. This relationship makes *r* a much more useful statistic than *d* when the goal of the analysis is prediction of an outcome (Costner, 1965).

One implication of this relationship is that even a dichotomous variable associated with a large *d* value may not be a particularly useful predictor when the base rates are very different. For example, Entry 6 of Table 3 suggests that individuals who are instructed to underreport pathology on the Minnesota Multiphasic Personality Inventory (MMPI) produce validity indicator scores that are on average about one standard deviation higher ($d = 0.94$) than those generated under normal instructions, a large effect. Suppose that in voluntary psychiatric settings only 2% of respondents have a vested interest in appearing overly healthy. Faking then turns out to have little relationship to respondents' scores ($r = .131$), even with the same standardized mean difference. Consequently, in the absence of information about the true base rate of underreporting, covarying or removing potentially invalid cases could result in unacceptably small improvements in scale validity (e.g., Piedmont, McCrae, Riemann, & Angleitner, 2000).

4. *r* is dependent on base rates, which has interpretive meaning in applied settings.

It has been suggested that the effect of base rate on the correlation coefficient can have interpretive value.

> Constraints on correlations associated with differences in distribution inherent in the constructs are not artifacts but have real interpretive meaning. . . . The observed correlation between smoking and lung cancer is about .10 . . . . There is no artifact of distribution here; even though the risk of cancer is about 11 times as high for smokers, the vast majority of both smokers and nonsmokers alike will not contract lung cancer, and the relationship is low because of the nonassociation in these many cases. (Cohen, Cohen, West, & Aiken, 2003, p. 54)

Similarly, as the proportion of individuals attempting to underreport on the MMPI declines, the proportion of false positive cases increases. r_{pb} changes to reflect this decline in predictive power. Consistent with the literature on maximizing the accuracy of diagnostic inferences (e.g., Meehl & Rosen, 1955), this makes *r* a more ecologically valid indicator of the effectiveness of the dichotomous variable as a predictor of the outcome than *d* when the true base rate is considered. This is a particularly valuable feature of the correlation coefficient as an indicator of the extent to which one variable can achieve practical utility as a predictor of another.

Even so, it would be a mistake to use the correlation coefficient as sufficient evidence of the relative importance of a risk factor. For example, more than half the American population is now considered overweight if not obese (Flegal, Carroll, Ogden, & Johnson, 2002). If the proportion of overweight adults continues to rise (diverging more from a base rate of .50), the correlation between being overweight and medical complications associated with excess weight in the general population will actually decline, even as weight continues to increase in importance as a risk factor.

Advantages of d Over r

1. Mean differences are particularly relevant for experimental or treatment effects.

Just as the nature of *r* makes it a more useful statistic when the goal is to determine the relationship between a predictor and a criterion, the nature of *d* makes it a more useful and readily understood statistic when the goal is simply to determine the amount of difference in the impact of two experimental conditions or treatments. However, as was noted in connection with the discussion of Figure 2, *d* is not the best indicator of the overall societal impact of an intervention if the population of individuals who receive the treatment is small relative to the total population.

2. *d* behaves more intuitively.

The sensitivity to base-rate differences can lead to some counterintuitive results for *r*. For example, suppose after preliminary analysis a researcher decides to increase the sample size as a means of increasing power. If the subsequent recruitment rate varies across groups and exacerbates a difference in base rates, the overall correlation can actually decline as a result of recruiting, though *d* does not. On the other hand, a decline in *r* can be used to warn the researcher that the sampling method is inefficient.

A second case of base-rate sensitivity producing unexpected results can occur when subgroups are combined. In a recent study (Blanchard, McGrath, Pogge, & Khadivi, 2003), college students completed the MMPI under instructions either to “fake bad” in a manner appropriate to mimic the results for someone not guilty by reason of insanity (forensic feigners) or to achieve psychiatric hospitalization (psychiatric feigners). These groups were then compared with psychiatric patients who completed the MMPI under standard instructions (see Table 4). When forensic feigners were compared with psychiatric patients on eight indicators of malingering, the mean *d* value was 1.98, whereas the mean *d* for comparing psychiatric feigners to psychiatric patients was 2.39. When both groups of feigners were combined in a composite analysis, the mean *d* value was 2.20, falling between the two subgroup means as one would expect.

Across the same eight predictors, the mean correlation between group membership and scale score was .39 for forensic feigners and .49 for psychiatric feigners. However, when the two groups were combined, so that the number of feigners was doubled, the base rate of feigners increased from .053 in the forensic condition and .061 in the psychihad increased), the mean correlation increased substantially to .54.[8]

3. *d* estimates effects independent of base rates.

A case may be made for base-rate-insensitive statistics as a general indicator of effect size when the base rate is subject to change across time and situation. Suppose the goal is to estimate the degree to which psychotherapy has been helpful for depression. If *r* is used to evaluate the relationship between treatment choice and ratings of improvement, the statistic will lose generalizability as the proportion of the population of depressives who have received treatment changes. In addition, to the extent that base rates fluctuate from sample to sample for nonsubstantive reasons when conducting a meta-analysis, one would expect greater confounding variability across studies in *r* (which responds to these nonsubstantive fluctuations), when compared with *d* (which does not).

As a result, *d* can provide a better estimate of the "transportability" of an effect to an alternative context where the base rates differ. For instance, parental susceptibility to stress may have a very small association, as measured by *r*, with the incidence of child physical abuse when studied in the general population where the incidence of abuse is quite low. These findings would suggest that interventions designed to bolster coping and stress resistance in parents may have little practical value for actually reducing abuse. However, if the same finding is accompanied by a relatively large *d* value, it would suggest that parental susceptibility to stress is nonetheless relatively important in the limited number of cases in which abuse actually occurs. As such, the *d* value accurately reveals that the stress–abuse relationship will become more apparent in settings in which the base rate for abuse is higher, suggesting, for example, that parental susceptibility to stress should be a more meaningful target of intervention for families in many clinical or forensic settings. As noted previously, the lack of sensitivity to base-rate change has by itself led some writers to prefer base-rate-insensitive statistics.

Choosing What to Report and How to Interpret the Effects

So both statistics have some desirable characteristics. How then is one to proceed? Some of the discussion suggests *r* is particularly suited for cases in which the task is to evaluate criterion-related validity. *d* is more appropriate when the goal is to determine the effect of an intervention or experimental manipulation. Furthermore, still other statistics may be more appropriate when the issue has to do with risk factors for negative outcomes. At times the distinction between these contexts may not be straightforward though. For example, though most of the studies that have evaluated the effectiveness of the MMPI as an indicator of faking good or faking bad have used experimental designs, these are analog studies of a prediction problem, and so *r* would typically be the more appropriate effect-size indicator assuming an ecologically valid estimate of the base rate is available. Similarly, even in experimental social research, in which *d* is the more commonly used effect size, the ultimate goal can be the prediction of real-world outcomes (e.g., Anderson et al., 1999; Funder & Ozer, 1983), a goal for which *r* is again defensibly the better measure. The preceding discussion leads us to the following recommendations, impact in real-world situations. A predictor that the correlation coefficient suggests is fairly weak can in fact prove to be quite powerful when considered in light of the inherent difficulty of predicting a rare phenomenon.

Several different approaches to the interpretation of the effect size can be suggested that take these multiple perspectives into account. For example, Rosenthal and Rubin (1982) recommended the binomial effect-size display as a general indicator of the true size of an effect regardless of the distributions of any dichotomous variables involved. However, this argument has been strongly criticized (e.g., Hsu, 2004).

Instead of relying on newer statistics, two options are available that use the familiar d and r_{pb} statistics. One option would involve the use of the standard fixed interpreindicate the d

of both *r* and *d* or through simultaneous comparison of one statistic to both standard and adjusted interpretive benchmarks. With regard to the latter possibility, we are reminded of the concerns Cohen (1988) raised with the introduction of his benchmarks.

> The terms “small,” “medium,” and “large” are relative not only to each other but to the area of behavioral science or even more particularly to the specific content and research method being employed . . . [T]here is a certain risk inherent in offering conventional operational definitions for these terms for use . . . in as diverse a field of inquiry as behavioral science. (p. 25)

Although some progress has been made in suggesting benchmarks that are appropriate to specific areas of behavioral investigation (e.g., Hemphill, 2003; Richard et al., 2003), the preceding discussion suggests that base rates also can be used to adjust benchmarks to the situation. At the same time, it is not our intention to suggest that all effects based on disparate base rates should be interpreted from multiple perspectives. The researcher should evaluate whether it is important to understand an effect independently of the base rates that hold in a particular setting, whether it is important to consider the impact of base rates on the potential for prediction, or both. Effect sizes cannot be understood in a vacuum, and researchers have an obligation to consider the context or contexts in which an effect is to be understood.

References

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect-size indices: Problems with a commonly recommended formula.* Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED433353)

American Psychological Association. (1994). *Publication manual of the American Psychological Association*

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality & Social Psychology, 44,* 107–112.

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

Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association, 86,* 1085–1111.

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3,* 339–353.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

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

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58,* 78–79.

Hsu, L. M. (2004). Biases of success rate differences shown in binomial effect-size displays. *Psychological Methods, 9,* 183–197.

Hunter, A. A. (1973). On the validity of measures of association: The nominal–nominal, two-by-two case. *American Journal of Sociology, 79,* 99–109.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods, 4,* 257–271.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lester, D. (1995). The concentration of neurotransmitter metabolites in the cerebrospinal fluid of suicidal individuals: A meta-analysis. *Pharmacopsychiatry, 28,* 45–50.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks CA: Sage.

McKenna, M. C., Zevon, M. A., Corn, B., & Rounds, J. (1999). Psychosocial factors and the development of breast cancer: A meta-analysis. *Health Psychology, 18,* 520–531.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., et al. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part I.* Washington, DC: American Psychological Association.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Narrow, W. E., Rae, D. S., Robins, L. N., & Regier, D. A. (2002). Revised prevalence estimates of mental disorders in the United States: Using a clinical significance criterion to reconcile 2 surveys' estimates. *Archives of General Psychiatry, 59,* 115–123.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78,* 582–593.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review & D (efH4ueo358vefH4uecial)Psycho D (efH.n.7(de-24.0054(efH&*

Lucia, M. S., Parnes, H. L., et al. (2004). Prevalence of prostate cancer among men with a prostate-specific antigen level less than or equal to 4.0 ng per milliliter. *New England Journal of Medicine, 350,* 2239–2246.

Torgersen, S., Kringlen, E., & Cramer, V. (2001). The prevalence