Review of General Psychology
2009, Vol. 13, No. 2, 130–136

© 2009 American Psychological Association
1089-2680/09/$12.00 DOI: 10.1037/a0015103

Is Psychological Research Really as Good as Medical Research? Effect Size Comparisons Between Psychology and Medicine

Christopher J. Ferguson
Texas A&M International University

Researchers have looked at comparisons between medical epidemiological research and psychological research using effect size *r* in an effort to compare relative effects. Often the outcomes of such efforts have demonstrated comparatively low effects for medical epidemiology research in comparison with effect sizes seen in psychology. The conclusion has often been that relatively small effects seen in psychology research are as strong as those found in important epidemiological medical research. The author suggests that many of the calculated effect sizes from medical epidemiological research on which this conclusion has been based are flawed. Specifically, rather than calculating effect sizes for treatment, many results have been for a Treatment Effect × Disease Effect interaction that was irrelevant to the main study hypothesis. A technique for developing a "hypothesis-relevant" effect size *r* is proposed.

Keywords: effect size (statistical), null hypothesis testing, statistical analysis, statistical significance, medical research

In 2001, Bushman and Anderson attempted to defend the effect sizes found relating media violence exposure to aggression by comparing these effect sizes with those seen in research linking smoking to lung cancer. Bushman and Anderson calculated the effect size for smoking and lung cancer as $r = .4$. This figure was based on a frequency table of data from the Wynder and Graham (1950) study of smoking effects. Table 1 is based on data from Wynder and Graham (1950) and demonstrates the relationship between smoking and lung cancer. Subsequently, a medical doctor and a statistician commented that Bushman and Anderson had miscalculated the effect size for smoking and lung cancer, and that the actual effect size is closer to $r = .9$ (Block & Crain, 2007). This figure of $r = .9$ would appear to be closer to the American Cancer Society's (2007) statement that 87% of lung cancer cases can be blamed specifically on smoking. Bushman and Anderson used the frequencies presented in Table 1 to calculate their effect size, whereas Block and Crain (2007) correlated smoking level by proportions of individuals developing lung cancer, as presented in Figure 1. Both calculations are easily verifiable, leaving us with the question of who (if anyone) is right. In the current article, I address issues such as these, examining common pitfalls in attempting to translate epidemiological research into effect size *r* and other effect sizes commonly used in psychology.

Calculating Effect Sizes From Medical Epidemiological Research

Upfront it is worth noting several things to put into perspective the difficulties of translating epidemiological research into effect sizes such as *r* or *d*. First, effect sizes such as *r* work ideally with continuous data, whereas much of the medical epidemiological literature, such as Wynder and Graham (1950), may use either binomial data or data that are reduced to nominal or ordinal information. Second, much of the medical literature refers to effect size estimates such as relative risk (RR) and odds ratios (Rosenthal & Rosnow, 2008), which are difficult to translate into effect sizes such as *r* and *d* (Rosenthal & DiMatteo, 2001). Third, in examining diseases that may occur uncommonly or rarely in the general population, epidemiological research may collect data on thousands of participants, from which only a small percentage of them (i.e., those who actually are exposed to a bacterium, or are actually at risk of developing an illness in the absence of a preventative intervention) is actually relevant to the study hypothesis. It is this third point that will occupy the central thesis of the current article, namely, that "hypothesis-irrelevant" cases in frequency counts have corrupted effect sizes calculated from medical epidemiological research. Put another way, effect sizes calculated from raw epidemiological frequency counts (e.g., Bushman & Anderson, 2001; Rosenthal, 1990; Rosnow & Rosenthal, 2003) confound the hypothesis-relevant treatment effect with the effect of the disease process itself on the general population.

To illustrate this point, consider the following hypothetical example. In the "Wild West," researchers Smith and Wesson are hired by the Dastardly Bandit Company (DBC) in response to the increased use of Gatling guns to protect stagecoaches. The DBC would like Smith and Wesson to develop a bulletproof vest to protect their bandits as they attempt to rob stagecoaches because, to date, casualties have been too high. Smith and Wesson begin by examining the incidence of casualties due to Gatling guns during stagecoach robberies. In Table 2, two groups are used (A and B), but these are essentially identical as neither have bulletproof vests. From Table 2 we can mainly see the effect of the Gatling guns on death, namely, that the Gatling guns are about 50% efficient. Bad news for bandits indeed! Note that this figure is difficult to communicate into an effect size such as *r* as the Gatling guns are a

Christopher J. Ferguson, Department of Behavioral, Applied Sciences and Criminal Justice, Texas A&M International University.

Correspondence concerning this article should be addressed to Christopher J. Ferguson, Department of Behavioral, Applied Sciences and Criminal Justice, Texas A&M International University, 5201 University Boulevard, Laredo, TX 78041. E-mail: CJFerguson1111@aol.com

Table 1
Frequency Table for Smoking and Lung Cancer

Smoking level (over 20 years)	No cancer (n)	Cancer (n)	Cancer probability (%)
None	114	8	7
1–9 cigarettes/day	90	14	13
10–15 cigarettes/day	148	61	29
16–20 cigarettes/day	278	213	43
21–34 cigarettes/day	90	187	68
35+ cigarettes/day	59	123	68

Table 2
Effects of Gatling Guns on Casualties in the Wild West

Participant group	Hit by Gatling gun	Not hit by Gatling gun
Group A (no bulletproof vest)	50	50
Group B (no bulletproof vest)	50	50

constant (all 200 participants are shot at, so Gatling gun as a variable has no variance). Nonetheless, we can see that this table does communicate the *effect* of the Gatling guns. Consider Table 3, which demonstrates what may have happened before the stagecoaches used Gatling guns and had to rely only on rifles. The rifles are clearly less efficient, only about 10% so. The differences between Table 2 and Table 3 demonstrate differences in effect for the two weapon systems on the population of bandits. In other words, irrespective of the introduction of an intervention or treatment, there is already a weapon (or disease in an epidemiological sense) effect in force in these frequency tables. In regards to medical epidemiology, this is equivalent to the effect of varying diseases on a given population. Some diseases are relatively common and spread easily (or are genetically common in the population) and thus have a large effect on morbidity; other diseases are rare and have a much less pronounced effect on morbidity. It has been indicated previously that such base-rate issues can have an influence on resultant effect size estimates (McGrath & Meyer, 2006), often attenuating effect size estimates considerably.

What Smith and Wesson are interested in examining is the effect size for bulletproof vests on death (assuming a penetrating bullet wound always kills) for individuals who are hit by the Gatling gun. Smith and Wesson are not interested in the effect of bulletproof vests on individuals who are *not* hit by the Gatling gun because this simply makes no sense (i.e., it is hypothesis irrelevant). However, Smith and Wesson do not know in advance which bandits are going to be hit by the Gatling gun, so they examine a larger sample of potential Gatling gun victims. However, not all of these individuals in the sample will be *hypothesis relevant* because the hypothesis is concerned only with those who actually get hit. In their study, half of the bandits are randomly assigned to wear bulletproof vests (Group A), whereas the rest (Group B) are given some form of "placebo" (perhaps a heavy wool vest). Table 4 presents the frequency table for their results. Two hundred bandits are sent to rob stagecoaches. As expected, half of those in Group B are killed by the Gatling guns, whereas only 25% of the bandits in Group A are killed by the Gatling guns. Using methods suggested by Rosenthal and Rosnow (2008), the chi-square statistic can be calculated, $\chi^2 = 13.33$, which, using the formula $[(\chi^2/N)^{1/2}]$, is equivalent to $r = .26$. Although this effect appears to be moderate in size, it is a miscalculated underestimate. The effect size is underestimated because, in the chi-square analysis, many irrelevant cases were allowed to remain. In other words, $r = .26$ is the effect size for the Gatling Gun × Bulletproof Vest interaction term for the general population of bandits. It is not the effect size for the effect of bulletproof vests on those who are hit by the Gatling guns. We know from both Table 2 and from Group B in Table 4 that approximately 50 bandits can be expected to be hit and killed by the Gatling guns; what Smith and Wesson want to know is how much the bulletproof vests improve on the natural state of affairs. To calculate the effect size for bulletproof vests, we must include only "hypothesis-relevant" cases, namely, those bandits who were targeted by the Gatling gun.

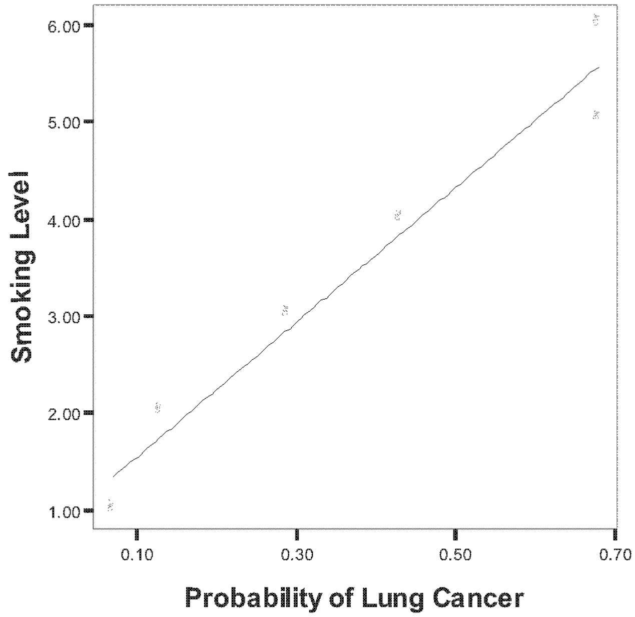

Figure 1. Smoking level and lung cancer probability.

Table 5 presents a "hypothesis-relevant" frequency table. This table removes the effect of the Gatling guns on the larger population from the equation, and focuses only on those individuals who we know are likely to have been targeted by the Gatling guns (i.e., hypothesis-relevant cases). Now the effectiveness of bulletproof vests over placebo in saving bandit lives can be calculated. Running chi-square tests on the resultant frequency table, we get $\chi^2 = 33.33$, which can be calculated as $r = .58$, which is significantly stronger than the previous calculated effect size of $r = .26$. This is the effect size that is relevant to the study hypothesis "How effective are

Table 3
Effects of Rifles on Casualties in the Wild West

Participant group	Hit by rifle	Not hit by rifle
Group A (no bulletproof vest)	50	500
Group B (no bulletproof vest)	50	500

Table 4
Effects of Gatling Guns and Bulletproof Vests on Casualties in the Wild West

Participant group	Dead	Alive
Group A (bulletproof vest)	25	75
Group B (no bulletproof vest)	50	50

bulletproof vests in preventing death for those who are hit by Gatling gun rounds?" Because the researchers Smith and Wesson had no idea in advance which bandits were going to be hit, they needed to use a "cast the net wide" sampling approach as represented in Table 4; however, calculating the effect size from this sample would have been an underestimate for the effectiveness of bulletproof vests. Unfortunately, I argue here, this is exactly what has been done in many cases in which authors have tried to translate medical epidemiological effect sizes to effect size *r*.

Part of the reason for the confusion I suspect lies in different sampling strategies used in most psychological treatment outcome research and experimental psychological research in contrast with those used in epidemiological research. For instance, in a traditional "treatment outcome" study in psychology, say examining a new treatment for depression, a sample would be recruited from among a population of individuals known to have a disorder that is relevant to the study hypothesis. Many treatment outcome studies begin by actively screening out individuals who do not possess a relevant disorder (e.g., Kunik et al., 2005). Thus, all individuals included in the final sample are "hypothesis relevant," and an effect size calculated on the results would be accurate. Including individuals without the relevant diagnosis into the effect size calculation would make little sense. I refer to this approach here as *targeted sampling*. In medical epidemiological research (and perhaps some psychological epidemiological research such as primary prevention research), the sampling technique is different. Because the focus is on *preventing* a disorder or studying etiological factors related to disorder development, epidemiological studies begin with broader samples of individuals, some who would develop the disorder and others who would not, irrespective of any preventative effort being examined. Because the investigators do not know in advance which participants are *hypothesis relevant,* data are collected on all participants whether or not they are hypothesis relevant. Only after the outcomes are known are individuals sorted and reflected in the relative risk assessments commonly employed in medical settings (Rosnow & Rosenthal, 2003). I refer to this as *net sampling*, quite literally from "cast the net wide." As mentioned earlier, effect size estimates used in medical research such as relative risk do not translate well into effect sizes such as *r* (Rosenthal & DiMatteo, 2001). In calculating effect size *r* from binomial or nominal × ordinal frequency tables, psychologists need to be careful to include only those cases that are *hypothesis relevant*, which may not be immediately apparent from the original data table. If this is not done, resultant effect sizes will be confounded by the base rate of the disease itself (Fleiss, Levin, & Paik, 2003; McGrath & Meyer, 2006). One approach used in the past (although seldom appropriately applied to psychology) to correct for this is the calculation of a "maximum phi" (equivalent to a maximum *r* value), which can be used to adjust the observed *r* (Guilford, 1965). This approach appears to be more of an estimate than a direct calculation, however, and may continue to attenuate actual effect sizes. The influence of sample size on *r* (or phi [ϕ]) is discussed below, and this "observed phi/maximum phi ratio" is arguably still influenced by sample size so long as sample size remains in the calculations. Use of the adjusted $\phi/\phi_{\max}$ is suggested only when naturalistic sampling has been employed (Carroll, 1961).

Table 5
Effects of Bulletproof Vests on Survival in the Wild West

Participant group	Dead	Alive
Group A (bulletproof vest)	25	25
Group B (no bulletproof vest)	50	0

Two of the pioneers regarding the conversion of medical epidemiological studies to effect sizes are Rosnow and Rosenthal (2003). It is worth noting that my comments here are not intended as criticism of their work, but rather as an attempt to refine and improve on their efforts given their considerable strides. Indeed, Rosnow and Rosenthal have elucidated most of the methods involved in converting medical effect sizes into *r* or *d*. I argue here that the only step remaining is detailing the development of appropriate hypothesis-relevant frequency tables. However, this is a critical step, without which medical effect sizes may appear to be artificially low. For instance, Rosnow and Rosenthal calculate the effect size for the Salk vaccine as $r = .011$. It would likely be fair to say that the hypothesis underlying the effectiveness of the Salk vaccine would be something along this line: "The Salk vaccine is effective in preventing polio in individuals who are exposed to the polio virus." In an experimental study, we might take targeted samples, give one group the vaccine, the other no vaccine, and deliberately expose all individuals to polio. In such a case, the effect size could be directly calculated because all individuals are hypothesis relevant (i.e., all have been exposed to polio). Naturally, this would be highly unethical (as well as illegal), and so investigators must rely on a net sampling approach that includes countless numbers of hypothesis-irrelevant cases.

The frequency table for this calculation (data from Rosnow & Rosenthal, 2003, p. 226, Table 2) is presented in Table 6. The first

Table 6
Salk Vaccine Frequency Tables; No Treatment Expected, Net Sampled, and Hypothesis Relevant

Condition	Polio present	Polio absent	Effect size
1. Disease only, no treatment			
No vaccine (Group 1)	115	200,712	
No vaccine (Group 2)	115	201,114	
			.000573%
2. Net sampled			
Vaccine	33	200,712	
No vaccine	115	201,114	
			$r = .011$
3. Hypothesis relevant			
Vaccine	33	82	
No vaccine	115	0	
			$r = .74$

Note. Frequency tables adjusted for sample size.

set of data presents the data that would be expected were vaccine simply unavailable. Both groups (with no variance between the groups because neither receives vaccine) would be expected to have about equal rates of polio virus (adjusted for sample size, although in this case the sample sizes are 99.8% similar). These data present the effect of the disease, which can be represented as a transmission rate of approximately .000573%. Thus, polio is rather rare in the general population, requiring net sampling. Some may protest that the disease is not an effect but rather an outcome, but this is incorrect. The disease is the action of a microbe, genetic inheritance, toxin exposure, and so forth (microbe in the case of polio), and is indeed an effect. The outcome is morbidity or death. Data Set 2 represents the original data sample, on which Rosnow and Rosenthal (2003) calculated a chi square of 45.25. From this, r is calculated by dividing the chi-square statistic by the sample size (401,974) and taking the square root of this result. Not surprisingly, any chi-square statistic divided by such a large sample is going to be miniscule, and correspondingly the effect size that Rosnow and Rosenthal calculate is $r = .011$, with r^2 of effectively zero. Rosnow and Rosenthal interpret this as evidence as support that relatively small effect sizes nonetheless may signify important results.

With respect for Rosnow and Rosenthal, I assert here that this conclusion is mistaken. An effect size r^2 that is zero means there is zero overlap in variance (or nearly so) between Salk vaccine administration and protection from polio. There essentially is no way to interpret this other than to conclude that the Salk vaccine is simply ineffective and should be seen to produce almost no change in the incidence of polio. Yet, we can see from the frequency table that administration of the Salk vaccine results in more than a threefold reduction in risk of polio (RR = 3.48). As such, clearly the Salk vaccine *is* effective. How could the effect size estimate be so off? As discussed in the example of the Gatling guns above, the reason is due to leaving the effect of the polio disease itself (which afflicts only .000573% of the net sample) in the equation. Thus, $r = .011$ is not the effect size for the Salk vaccine; it is the combined effect sizes for the Salk vaccine and the polio virus itself, which effectively drowns out the effect of the Salk vaccine.

Data Set 3 corrects for this by removing all hypothesis-irrelevant cases (cases that result from net rather than targeted sampling). Because we expected 115 disease cases in each group without treatment, we can now limit the frequency table to only those cases, and see what the actual impact of the Salk vaccine is on preventing polio in individuals who are exposed to polio (99.999427% of this net sample either was not exposed to polio or had natural immunity; either way, they are irrelevant cases in regard to the effectiveness of the Salk vaccine). If the Salk vaccine is ineffective, we could have expected 115 cases of polio in the vaccination group, just as with the no-vaccination group. Thus, we can limit our analyses to those cases only (115 vaccine, 115 no vaccine). However, we have only 33 polio cases in the vaccine group, leaving 82 of our expected 115 as "helped" by the Salk vaccine. Calculating chi square on this hypothesis-relevant frequency table, we get the result $\chi^2 = 127.43$, which corresponds to $r = .74$. Even intuitively, this would appear to be a much better representation of the effect size of the Salk vaccine, better corresponding to a relative risk of 3.48 (although again, r and RR are not easily translatable). From this, we can see that an intervention that had been successful in preventing 100% of the expected incidence of disease would demonstrate $r = 1.00$ as its effect was, in essence, perfect. In the event that a treatment was *worse* than the disease, the effect size would be calculated in a similar manner (with the lower functioning "treatment" becoming the baseline) but with a negative sign to indicate the direction of effect.

Given that many authors use these calculated effect sizes from medical epidemiological research to defend effect sizes seen in psychology (e.g., Bem & Honorton, 1994; Bushman & Anderson, 2001; Rosnow & Rosenthal, 2003), this is no small issue. I argue here that these comparisons have been in error and that the effect sizes of medical epidemiological research are, in truth, much larger than have been reported. Indeed, other authors have already commented on the mathematical problems underlying these effect size estimates (Hsu, 2004; Kraemer, 2006). Arguably, such comparisons were always questionable, effectively comparing apples to oranges. The outcome variables in medical epidemiological research often benefit from "perfect" or "near perfect" validity. To put it bluntly, no one need worry whether death is a valid measure of death. The same cannot be said for psychological outcome variables, either self-report or behavioral, where validity often remains a considerable problem (Tedeschi & Quigley, 2000).

To highlight this issue, I examined the effect sizes calculated for a variety of research results presented by Rosnow and Rosenthal (2003) that included medical epidemiological results, psychological meta-analyses, and targeted sampling psychological research. In the original article, these were presented in Table 3 (p. 227). In this article, Table 7 presents these studies, their sample sizes, the original effect sizes (Rosnow & Rosenthal), represented as $r_{t\times d}$ (Treatment × Disease combined effect size), the corrected hypothesis-relevant effect sizes represented as r_h, and r^2 for the hypothesis-relevant effect sizes. Despite helpful efforts of one author (Rosenthal) of the Rosnow and Rosenthal article and an attempt to contact the publication source (i.e., *Harvard Gazette*) for one manuscript on cisplatin and vinblastine (Cromie, 1990), this article could not be located. Attempts to identify the primary source were likewise unsuccessful and as such this example was removed from the analysis.

Unlike statistical significance, effect size is not (or should not be) influenced by sample size (Chow, 1988). When a body of research appears to demonstrate a relationship between effect size and sample size, this is often an indication of problems with this body of research (Ferguson, 2007). As can be seen from looking at Table 7, in the body of data presented by Rosnow and Rosenthal (2003), larger sample sizes are associated with smaller effects. This is likely because the larger sample sizes are using larger "net" samples to capture relatively small numbers of hypothesis-relevant cases. As the method of calculating r uses sample size (square root of the chi-square statistic divided by sample size), the result is artificially deflated r values. This can be represented quantitatively. Because neither sample size nor effect size is normally distributed (Shapiro–Wilk $p < .01$ for both), I used Spearman-rho correlations to examine the relationship between effect size and sample size on the studies in Table 7. Results indicated a significant relationship between sample size and the effect sizes $r_{t\times d}$ calculated by Rosnow and Rosenthal, $\rho = -.89$, $p < .01$. Put another way, 78% of the variance in the effect sizes calculated by Rosnow and Rosenthal can be attributed to the sample size of the study, with larger samples producing smaller calculated effect sizes. This is clearly an indication of a problem in the methodology

Table 7
Effect Sizes Seen in Medical Epidemiological, Meta-Analytic, and Targeted Sampling Results Using Net Sampled and Hypothesis-Relevant Approaches

Independent variable	Dependent variable	n	$r_{t\times d}$	r_h	r^2
Salk vaccine (Francis et al., 1955)	Polio contraction	401,826	.011	.74	.55
Aspirin (Steering Committee of the Physicians Health Study Research Group, 1988)	Heart attacks	22,071	.03	.52	.27
Beta carotene (Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group, 1994)	Death	29,133	.03	.19	.04
Streptokinase (Gruppo Italiano per lo Studio della Streptochinasi Nell'Infarto Miocardico, 1986)	Death	11,806	.03	.31	.09
Propranolol (Kolata, 1981)	Death	3,837	.04	.39	.15
Magnesium (Foreman, 1995)	Convulsions	2138	.07	1.00	1.00
Vietnam veteran status (Centers for Disease Control Vietnam Experience Study, 1988)	Alcohol problems	4,462	.07	.44	.20
Garlic (Goldfinger, 1991)	Death	432	.09	.58	.33
Indinavir (Knox, 1997)	AIDS events	1,156	.09	.56	.31
Testosterone (Dabbs & Morris, 1990)	Adult delinquency	4,462	.12	.62	.38
Hosp. vs. tx. choice (Walsh et al., 1991)	Alcohol problems	144	.13	.51	.26
Cyclosporine (Canadian Multicentre Transplant Study Group, 1983)	Death	209	.15	.75	.56
Warfarin (Grady, 2003)	Blood clots	508	.15	.67	.45
ESP (Bem & Honorton, 1994)	Accuracy	21.8[a]	.16	.16	.03
AZT for neonates (Altman, 1994)	HIV infection	364	.21	.71	.51
Cholesterol treatment (Roberts, 1987)	Coronary status	162	.22	.46	.21
AZT (Barnes, 1986)	Death	282	.23	.94	.88
Tx. choice vs. AA (Walsh et al., 1991)	Alcohol problems	154	.27	.49	.24
Psychotherapy (Smith, Glass, & Miller, 1980)	Mental health	111[b]	.39	.39	.15
Hosp. vs. AA (Walsh et al., 1991)	Alcohol problems	156	.40	.68	.46
Anxiety (Rosnow, 1991)	Rumor mongering	84.38[a]	.48	.48	.23
Progesterone (Contraceptive trials, 1996)	HIV infection	28	.65	.87	.75

Note. Hosp. = compulsory hospitalization; tx. = treatment; AA = Alcoholics Anonymous.
[a] Average sample size for studies included in meta-analysis. [b] Estimated from imprecise figure of "tens of thousands" (Smith et al., 1980). The mid figure (50,000) was used and divided by number of studies (450) in this meta-analysis.

used to calculate effect size $r_{t\times d}$ using the raw frequency data tables from these studies. By contrast, the hypothesis-relevant effect sizes r_h calculated using the methods discussed in this article were not significantly related to sample size, $\rho = .19$, $p > .05$ (95% CI = $-.25 < \rho < .57$). It is thus concluded here that r_h may be a more accurate calculation of effect sizes seen in medical epidemiological research than is $r_{t\times d}$ as used in Rosnow and Rosenthal, although ultimately this depends on the type of question that the researcher is hoping to answer.

The Pitfalls of Effect Size Calculation

Returning to the debate between Bushman and Anderson (2001) and Block and Crain (2007) regarding the effect size of smoking on lung cancer, we can see that translating medical epidemiological results to effect sizes such as r is not as easy as may have been hoped. Given that Bushman and Anderson used a net sample frequency table (see Table 1) in calculating their results, their figure of $r = .40$ is problematic in the same ways in which the results of Rosnow and Rosenthal (2003) were problematic. It is problematic in that it also presents an average effect size across all levels of smoking, rather than a comparison of nonsmokers to two-pack-a-day smokers. Naturally, the number of cigarettes consumed daily would influence the effect on lung cancer. As such, we can say with certainty that the estimate of $r = .40$ is wrong. The figure reached by Block and Crain of $r = .90$ does agree with data from the American Cancer Society (2007), which lists a relative risk (RR = 23.3) and an estimated 87% variance overlap between smoking and lung cancer. However, using the r_h method described in this article, the effect size for smoking is found to be $r = .54$, higher than that suggested by Bushman and Anderson, yet lower than would have been expected from the American Cancer Society data. Why do we still find a discrepancy?

There are several possible reasons. The first is simply that conversions of nominal and ordinal data into effect size r may inherently risk some attenuation of the estimated effect size due to reduced variance. This may be somewhat less of a problem where outcomes are naturally dichotomous (e.g., live vs. die), and more so a problem when false dichotomies or false ordinal variables are created. This is the case with the smoking data wherein data that were effectively ratio in scale (number of cigarettes smoked) were reduced to an arbitrary ordinal scale. Also at the high end of smoking behavior, there appears to be little difference between 21 through 34 cigarettes/day and 35+ cigarettes/day. This lack of variance at the top may effectively be a ceiling effect, reducing the effect size estimate, despite the fact that cigarettes actually have a high impact on lung cancer rates. The conclusion that one reaches is that even with the revised r_h method for calculating effect size r from epidemiological research, such calculations may continue to be prone to underestimation. With that in mind, I make the following observations and suggestions.

1. Ideally, the best option would be for psychologists to begin to familiarize themselves with effect size estimates used in medical research such as relative risk, risk difference, and odds ratio. Although Rosnow and Rosenthal (2003) argue that r is superior to relative risk as an effect size estimate, this is likely to depend on sampling strategies used (naturalistic, random sampling, etc.).

2. With studies that use "targeted" sampling (e.g., randomized treatment outcome, most "basic" experimental studies), the $r_{t \times d}$ methods of effect size calculation pioneered by Rosnow and Rosenthal (2003) are likely to be adequate, although again this will depend on specific sampling strategies. For epidemiological studies (medical or psychological), if it is necessary to calculate r, the r_h method should be used to assure that the effect size is hypothesis relevant and not attenuated by the disease base rate (see Crow, 1991; McGraw, 1991).

3. Comparisons between medical research and psychological research should simply be avoided. Much of medical research benefits from variables with "perfect" validity (death, for instance), whereas I argue here that issues of measuring error (i.e., reliability, validity) are a greater concern for psychology than for much of medical research. Certainly, medical science is not immune to validity issues, and the well-versed reader is likely to be able to provide examples in which medical researchers have experienced measurement problems with medical disorders. Yet, the comparisons made in much of the literature between medical science and psychological science (Bem & Honorton, 1994; Bushman & Anderson, 2001; Rosnow & Rosenthal, 2003) have involved rather conclusive medical outcomes such as death or significant morbidity. This is perhaps because several of these studies (i.e., the Salk vaccine, the aspirin–heart disease link) are perceived by the general public as "conclusive." I submit here that the reason why such comparisons are popular (particularly using the effect size-attenuating $r_{t \times d}$ method for medical epidemiological research) is to assuage psychologists' sense of insecurity at being perceived as a "soft" science while medical research is perceived as a "hard" science. Of course, the popularity of these comparisons has rested on the belief that medical effect sizes were smaller than those found in psychology. As can be seen from Table 7, however, medical effect sizes are generally of greater size than those found in psychology. My suggestion is, effectively, for psychologists to "make their peace" with this reality and to remember that they probably knew this going into the field. In fact, one could argue that psychology is a more nuanced, subtle, and complex field given the difficulties inherent in examining the human mind.

4. I further submit that psychologists, in general, have followed the recommendations suggested by the American Psychological Association Task Force on Statistical Inference (Wilkinson & Task Force on Statistical Inference, 1999) by the letter, but have failed in spirit. I can report only on my subjective perception here, but to the extent that effect sizes are reported in social science literature (and this remains imperfect), they are seldom interpreted. When they are interpreted, as noted in the examples above, it is at times with the vein of suggesting that any effect size including those with a coefficient of determination of zero may retain practical significance. It seems that confidence intervals around effect sizes almost never are reported. Although Cohen (1994) pointed out the difficulties in setting standards for the interpretation of effect size, I would suggest that failing to have any at all renders the reporting of effect size inherently impotent and functionless. Perhaps, as Cohen noted, any standards would be arbitrary, but then again so is the $p < .05$ standard. Indeed, I would suggest that the effect size interpretations provided by Cohen are themselves too generous in light of the data from medical studies and other sciences. Naturally, the interpretation of effect sizes will be dependent on many factors, such as the costs and benefits of a causal effect. Further discussion within the field of psychology is warranted. Perhaps, however, it may useful also to honestly compare psychological outcomes with those found in other disciplines such as medicine or, indeed, physics, so that our results may be more properly put into perspective.

References

Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine, 330,* 1029–1035.

Altman, L. K. (1994, February 21). In major finding, drug limits H. I. V. infection in newborns. *New York Times,* pp. A1–A13.

American Cancer Society. (2007). *Smoking and cancer mortality table.* Retrieved October 2, 2007, from http://www.cancer.org/docroot/PED/content/PED_10_2X_Smoking_and_Cancer_Mortality_Table.asp

Barnes, D. M. (1986, October 3). Promising results halt trial of anti-AIDS drug. *Science, 234,* 15–16.

Bem, D., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin, 115,* 4–18.

Block, J., & Crain, B. (2007). Omissions and errors in "Media Violence and the American Public." *American Psychologist, 62,* 252–253.

Bushman, B., & Anderson, C. (2001). Media violence and the American public. *American Psychologist, 56,* 477–489.

Canadian Multicentre Transplant Study Group. (1983). A randomized clinical trial of cyclosporine in cadaveric renal transplantation. *New England Journal of Medicine, 309,* 809–815.

Carroll, J. B. (1961). The nature of data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–372.

Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: 1. Psychosocial characteristics. *Journal of the American Medical Association, 259,* 2701–2707.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103,* 105–110.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Contraceptive trials set for a link to AIDS research. (1996, May 7). *Boston Globe,* p. B1.

Cromie, W. J. (1990, October 5). Report: Drugs affect lung cancer survival. *Harvard Gazette, 1,* 10.

Cromie, W. J. (1991, September 13). Study: Hospitalization recommended for problem drinkers. *Harvard Gazette,* 3–4.

Crow, E. (1991). Response to Rosenthal's comment, "How Are We Doing in Soft Psychology?" *American Psychologist, 46,* 1083.

Dabbs, J. M., Jr., & Morris, R. (1990). Testosterone, social class, and antisocial behavior in a sample of 4,462 men. *Psychological Science, 1,* 209–211.

Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior, 12,* 470–482.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York: Wiley.

Foreman, J. (1995, July 27). Medical notebook: A new confirmation for a pregnancy drug. *Boston Globe,* p. B3.

Francis, T., Korns, R., Voight, R., Boisen, M., Hemphill, F., Napier, J., & Tolchinsky, E. (1955). An evaluation of the 1954 poliomyelitis vaccine trials—Summary report. *American Journal of Public Health, 45,* 1–63.

Goldfinger, S. E. (1991, August). Garlic: Good for what ails you. *Harvard Health Letter, 16*(10), 1–2.

Grady, D. (2003, February 25). Safe therapy is found for high blood-clot risk. *New York Times,* pp. A1–A22.

Gruppo Italiano per lo Studio della Streptochinasi Nell'Infarto Miocardico. (1986, February 22). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet, 1,* 397–402.

Guilford, J. P. (1965). The minimal phi coefficient and the maximal phi. *Educational and Psychological Measurement, 25,* 3–8.

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

Knox, R. A. (1997, February 25). AIDS trial terminated: 3-drug therapy hailed. *Boston Globe,* pp. A1–A16.

Kolata, G. B. (1981, November 13). Drug found to help heart attack survivors. *Science, 214,* 774–775.

Kraemer, H. C. (2006). A simple effect size indicator for two-group comparisons? A comment on $r_{\text{equivalent}}$. *Psychological Methods, 10,* 413–419.

Kunik, M., Roundy, K., Veazey, C., Souchek, J., Richardson, P., Densmore, D., et al. (2005). Surprisingly high prevalence of anxiety and depression in chronic breathing disorders. *Chest, 127,* 1205–1211.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11,* 386–401.

McGraw, K. (1991). Problems with BESD: A comment on Rosenthal's "How Are We Doing in Soft Psychology?" *American Psychologist, 46,* 1084–1086.

Roberts, L. (1987, July 3). Study bolsters case against cholesterol. *Science, 237,* 28–29.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775–777.

Rosenthal, R., & DiMatteo, M. (2001). Meta analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

Rosenthal, R., & Rosnow, R. (2008). *Essentials of behavioral research* (3rd ed.). New York: McGraw-Hill.

Rosnow, R. L. (1991). Inside rumor: A personal journey. *American Psychologist, 46,* 484–496.

Rosnow, R., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57,* 221–237.

Smith, M., Glass, G., & Miller, T. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 318,* 261–264.

Tedeschi, J., & Quigley, B. (2000). A further comment on the construct validity of laboratory aggression paradigms: A response to Giancola and Chermack. *Aggression and Violent Behavior, 5,* 127–136.

Walsh, D., Hingson, R., Merrigan, D. Levenson, S., Cupples, L., Heeren, T., et al. (1991). A randomized trial of treatment options for alcohol-abusing workers. *New England Journal of Medicine, 325,* 775–782.

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

Wynder, F., & Graham, E. (1950). Tobacco smoking as a possible etiological factor in brochiogenic carcinoma. *Journal of the American Medical Association, 143,* 329–336.

Received May 14, 2008
Revision received November 2, 2008
Accepted November 12, 2008 ■