CORRESPONDENCE

Why Vote-Count Reviews Don’t Count

To the Editor:

In general, there are two common approaches to a quantitative literature review: some form of meta-analysis or a vote count. The vote-count procedure is typically employed as follows: create two stacks of papers—stack one contains those papers that are statistically significant and consistent with a hypothesis and stack two contains papers that are not statistically significant as well as papers that are statistically significant and not consistent with a hypothesis. If stack one is greater than stack two, we conclude that there is an effect. Otherwise we conclude that there is no effect.

McCaffery et al (1999) consider the pros and cons of conducting a meta-analysis of structural magnetic resonance imaging studies in patients with schizophrenia and conclude that a meta-analysis is not appropriate. I do not fault them for this—reasonable people can differ about the appropriate application of meta-analytic techniques in any particular case. However, McCaffery and colleagues, like most investigators, fail to consider the pros and cons of a vote-count review. The vote-count procedure is a venerable one for a general discussion of the vote-count procedure, see Bushman (1994), but there are important limitations to the method.

The argument I am about to make is not original. It has been made most convincingly by Hedges and Olkin (1980, 1985, 48–52). But it bears repeating, since it seems that few investigators are aware of its crucial implications for our field.

Let me begin by defining some terms: effect size is a standardized mean difference—the difference between two means expressed in SD units. An effect size of .3 indicates that the means of the two groups in question are 0.3 SD units apart. Statistical power is defined as the probability, given a certain true effect size, that a study with a given sample size will yield a statistically significant result (herein, we will assume a one-tailed test and an α = .05). In planning a study, one typically aims to have a power of .8. In other words, with a true population effect size of X, and a sample size N, eight times out of 10 a study will yield a statistically significant result. But here is the rub: often the power of a particular study is less than .5. In other words, given that there is a true population effect, a study with a particular sample is more likely than not to fail to find statistically significant result.

To illustrate how low-power studies might impact a vote-count review, I propose the following thought experiment. Suppose we are God and therefore know that the true population effect size for the reduction in the volume of structure Z in patients with schizophrenia is .4. Now suppose that determining this effect size is of major importance to the human race and therefore 1 million studies, each with a sample of 30 patients with schizophrenia and 30 control subjects, are performed. Now suppose that someone decides to do a vote-count review of these 1 million studies. Surely, with 1 million studies, a vote-count review will come to the correct conclusion. Unfortunately, the probability that the vote-count review will come to the correct conclusion (that there is a reduction in the volume of Z) is nearly 0 (certainly p < .00000000000000000001). This is because the power of the average study is .45. Roughly 45% of the studies will find a statistically significant result, but 55% of the studies will not find a statistically significant result.

Of course, if one is certain that the studies to be reviewed have power greater than .5, then it may be reasonable to expect that a vote-count review will come to the correct conclusion. So, let us look at the meta-analyses of brain imaging studies in schizophrenia to see what the average power of the studies is. The meta-analysis by Ward et al. (1996) found a composite effect size for brain size reduction in schizophrenia to be .31, and there were approximately 28 subjects per group per study. (Wright et al [2000] independently arrive at a similar effect size—i.e., .25.) The average power of a study with these parameters is .31. (It is a coincidence in this case that the average power is the same value as the effect size.) Or, take studies of hippocampal size reduction in schizophrenia. The meta-analysis by Nelson et al. (1998) found a composite effect size of .39 for the reduction in size of the right hippocampus, and the average sample size per group in the reviewed studies was approximately 26. The average power of these studies is .40. My colleague, Lisa Konick, and I have just completed a meta-analysis of thalamic size reduction in schizophrenia and find an effect size of .30, with an average sample size of approximately 34 subjects per group per study (Konick and Friedman, in press). The average power of a study with these parameters is .34. A vote-count review of 1 million studies is certain to come to the wrong conclusion in all of these cases. Smaller vote-count reviews might find the correct answer by chance. But as the number of studies reviewed increases, the probability of finding the correct answer diminishes toward 0.

If one does not trust the results of meta-analyses in this field, then one can distrust the particular effect size numbers in the previous paragraph. But the effect size is some number. And if the effect size in question is in the small (.2) to moderate (.5) range, and the sample sizes per study per group remain near 30 or less, there is a strong likelihood that a vote-count review, regardless of how many studies are reviewed, will not yield the correct answer.

In summary, a vote-count review is likely to yield the wrong conclusion if most studies in a particular area of research have power less than .5. And, unfortunately, in structural imaging studies of patients with schizophrenia, many published studies, perhaps most, have power less than .5.

Lee Friedman

University Hospitals of Cleveland
Department of Psychiatry
Hanna Pavilion, B68
11100 Euclid Avenue
Cleveland OH 44106

© 2001 Society of Biological Psychiatry 0006-3223(00)01075-1
References


Reply

To the Editor:

We thank Dr. Friedman for his interest in our review of magnetic resonance imaging findings in schizophrenia. We also agree with his point that “reasonable people can differ about the appropriate application of meta-analytic techniques,” and he does not quibble about our not using meta-analysis. More specifically, we noted in our review (McCarley et al 1999, 1100) that for a review of this broad a scope, covering more than a decade and all published studies, meta-analysis was not the most appropriate approach, as it would yield a false sense of numerical exactness. We suggest that one can not simultaneously do a comprehensive review of a decade of studies, the goal of this review, and also perform a valid meta-analysis, unless the variables [of technical quality, ROI definition, and subject moderator variables] have remained constant, a constancy for which there is little evidence [italics added].

Where Dr. Friedman does have a problem with our review is in his reading our review as endorsing a simple-minded presentation of vote counting; in his reading we “fail to consider the pros and cons.” We italicize his to emphasize that, in fact, the review does note the pros and cons of this approach. Moreover, we used a method that, had Dr. Friedman read the relevant parts of the review, would have obviated his concern of a less than 10^{-20} probability of a correct conclusion, given an effect size of .4 and 10^5 studies in his example.

We call the reader’s attention to the section of our review on page 1100, beginning with: “Second, we need to comment on the tabular presentation of the studies in Table 1.” (We go on to note the need to provide literature references to back up review “expert opinions” so that the reader can judge the evidence. We then take up the problems of using meta-analysis in a comprehensive review of this nature. We encourage the reader to reread this section, since space prevents our repeating the details of argument here, and since Friedman did not contest this point.)

The key points that Dr. Friedman apparently missed are on the same page:

We, nonetheless, are sympathetic to readers who would like some more quantitative information and for this reason we have provided a table of subject N for each study in a summary table. It will be clear that the statistical power of negative studies with small subject N is less, and that their results consequently are less convincing than those negative studies with a larger N. We also, following a suggestion of Rosenthal (1987), computed the probability of the observed positive and negative statistical findings, using a two-tailed probability, and the alpha level of \( p < .05 \) of the studies. The reader will note that this procedure, like the more standard meta-analyses, assumes comparability of the studies with respect to measurements and subjects, although it does not assume normality of the distributions. We use this probability statistic with caution, since, in our opinion, it is hazardous to assume that the studies are comparable in methodology and subject characteristics; however, it may be of interest to the reader that, if one does assume comparability, the binomial theorem computation (using \( p < .05 \) for a positive study) shows that all ROIs surveyed in Table 1 show a two-tailed \( p < .05 \) for the number of positive studies, except for the fourth ventricle and cerebellum, and all ROIs are less than or equal to .002 except for the occipital lobe (.004).

Although this procedure has the same liabilities as meta-analysis (of which it is a nonparametric variant) in comparing possibly noncomparable studies, it is useful, we believe, because it gives the reader another point to figure into his/her weighting of the data and, of course, would avoid the conclusion Friedman introduces in his example. What he labels as a simple vote count was, in fact, not.

Let us go through the computations. The probability of a statistically significant result in a single study is \( p < .05 \). The probability of 10 studies each showing statistically significant results is \( p < .05 \times 10^{-2} = p < .00000000000009765625, or 9.765625 \times 10^{-14} \).

The reader and Dr. Friedman will recall that, for the case of a series of mixed successes and failures, the binomial theorem provides the correct computation of the probability for \( p \) successes and \( q \) failures when the true probability of a success is \( < .05 \), and this is the way we calculated the probabilities cited above for the summary in Table 1.

Thus, using this formula (Z approximation, since \( N \) is large) for Dr. Friedman’s example, the calculated probability of concluding there was no significant effect is \( < 10^{-15} \), a probability level that would give the reader a clear indication of which way the literature was pointing. Dr. Friedman incorrectly calculated that the probability of correctly finding a significant effect was...