under investigation, has been supported. Conversely, if the first clinical trial produces statistically significant results and the second investigation statistically nonsignificant results, then the usual conclusion is that the results of the two trials conflict. That is, there is a lack of consensus or a failure to replicate (table 2). The White and Black example, discussed earlier, illustrates that the presence of statistical disagreement does not mean there is a lack of support for the initial treatment effect.

Reliance on the results of statistical significance testing to confirm or refute empirical consensus has a long history in the social and behavioral sciences.[11,30-32] The practice of using the results of tests of statistical significance to confirm the success of replication efforts is referred to by Light and Smith[33] as the vote counting method. Hedges and Olkin[34] note that the "conventional vote counting or box score methodology uses the outcome of the tests of significance in a series of replicated studies to draw conclusions about the magnitude of the treatment effects." Using the vote counting method, the investigator identifies replicated trials and counts the number with statistically significant results and compares this number with the number of trials with statistically nonsignificant results. Vote counting fits the model of replication presented in table 2. Its main attraction is simplicity and the ability to quickly confirm consistent findings.

The logical and quantitative limitations of vote counting as a method of assessing successful replication efforts have been described in detail by Hedges and Olkin.[34] They note that with 10 trials, a sample size of 30 per trial, and a population effect size (d-index) of 0.50, a vote count will fail to detect the substantial treatment effect (d = 0.50) more than 90% of the time. In the behavioral and rehabilitation sciences, where samples are often small and effect sizes are in the range defined by Cohen[21] as small or medium, such a low probability of confirming replications and establishing a consistent treatment effect is not acceptable. A second, and perhaps more important problem with vote counting and the logic of replication presented in table 2 is that it ignores statistical power. When sample size and effect size are small, a simple vote count will often fail to identify a significant overall treatment effect and lead to the conclusion that a failure to replicate has occurred.

When two quantitative clinical trials are conducted, in which the second is a direct replication of the first, what is the probability that both investigations will produce statistically significant results? Based on the information contained in table 2 some investigators imagine that the probability of replication for the second experimental trial is 95%. This is not the case. Assuming that the null hypothesis is being tested at the $p < .05$ level of statistical significance, the probability that the null hypothesis will be accepted twice in succession when using conventional statistical methods is 0.95^2 or 0.9025. In contrast, the probability that it will be falsely rejected twice is 0.05^2 or 0.025. Thus, the overall probability of replication is 0.9025 + 0.025 or approximately 0.93. The value of 0.93 is an approximation that does not take into consideration the influence of statistical power. It is also important to note that 0.93 is the probability associated with replication before conducting the original investigation. If the investigator conducting the replication study is examining a genuine treatment effect, then statistical power becomes a critical factor in determining the success of any replication effort.

Assume a researcher has conducted a series of well-controlled replication trials with a mean power of 0.40, then the probability of replication, p (replication) = $0.40^2 + (1 - 0.40)^2$ is 0.52. This value is even less encouraging if we consider that because of the low power (0.40) the majority of the replications in this example represent confirmations of a false impression. If the power is 0.40, then the corresponding type 2 error rate is 0.60 or 60%. In fact, $(1 - 0.40)^2 + [0.40^2 + (1 - 0.40)^2] = 0.880$, or 88% of the replications may strengthen a false impression (a type 2 error).

As an illustration of the impact of statistical power on replication, we can further examine the investigations by Black and White described previously. Inspection of the two trials show that the power for White's trial is 0.60 (type 2 error rate = 0.40). The power for Black's trial is 0.18 (type 2 error = .82). Given a medium sized population effect (d-index = 0.50) there are only 11 chances in 100 ($0.60 \times 0.18 = 0.11$) that both trials would reject the null hypothesis. The odds are three times greater ($0.40 \times 0.82 = 0.33$) that neither trial would reject the null hypothesis. That is, it is three times more likely that there will be a successful replication of a negative (statistically nonsignificant) result, than of a positive (statistically significant) result. This paradoxical situation is the result of low statistical power.

IMPLICATIONS AND RECOMMENDATIONS

One consequence of a failure to replicate a previous clinical trial is a proliferation of investigations developed around hypothesized explanations for the statistical conflict. These explanations are usually based on conjecture related to the original theory or some flaw in the treatment technique. Rarely is the mundane hypothesis of mere sampling variation and/or low statistical power considered as the preferred explanation for the failure to reject the null hypothesis. Disagreements in the literature are more commonly used as springboards for variations of the original investigations. Exact replication in clinical research is rare.[35] Even rarer is replication with a more powerful design. Thus, the consequence of low power and statistical conclusion invalidity is a contradictory research literature that fails to establish consensus.

A second negative consequence of a failure to establish consensus is a disaffection with traditional quantitative methods. The apparent inability of experimental designs to produce statistically positive or consistent results may lead some researchers to abandon quantitative methods and search for alternative research strategies. The problems of research quality and inconsistent results associated with quantitative clinical trials are often a result of misinterpretation, rather than a deficiency of the techniques themselves. As illustrated previously, a failure to appreciate statistical conclusion invalidity can result in misinterpretation at several different levels.

Rehabilitation researchers can reduce statistical confusion