level, with a sample size of 20 and an effect size (d-index of 0.50), we find the tabled statistical power for Black's test is 0.18.

Earlier it was noted that power = 1 − type 2 error. A simple algebraic manipulation shows that type 2 error = 1 − power. The type 2 error rate for Black's study is (1 − 0.18) or 0.82. This means that there was an 82% chance that Black's decision (based on the statistically nonsignificant result) to not reject the null hypothesis was a mistake; a type 2 error.

To summarize, White conducted a clinical trial and reported a statistically significant result (rejected the null hypothesis). Black performed a replication trial and reported a statistically nonsignificant result (failed to reject the null hypothesis). When the results from the two investigations were analyzed to determine the effect size, both studies showed the same d-index (0.50). If the d-index, representing treatment impact, was the same for both studies, why did the statistical results disagree? The probability of committing an experimental error (type 1) in the White trial was 5%. The probability of an error (type 2) in Black's trial was 82%. The probability of a type 2 error in Black's clinical trial was high because of the low statistical power. The statistical power was low as a result of the small sample size.

It is important to understand that the researcher (or reader) never knows whether a type 1 or type 2 error has occurred in a given clinical trial. What the researcher (or reader) knows is the probability of an error occurring. In theory both Black and White could commit a type 1 and type 2 error. After the data are collected, analyzed and a decision made regarding the null hypothesis, White is restricted to making the correct decision or a type 1 error; and, Black to making a correct decision or a type 2 error. In this case, the high probability of a type 2 error in Black's trial (82%) is the most logical explanation for the discrepancy between the statistical results from the two investigations. The probability is very high that Black committed a type 2 error, ie, he failed to reject the null hypothesis when it should have been rejected.

The problem of low statistical power and type 2 errors in clinical research is not widely appreciated or understood. The focus in traditional quantitative clinical trials has been on the control and prevention of type 1 errors. The probability of committing a type 1 error is usually held at 5% ($p < .05$) or lower and little attention is devoted to the impact of type 2 errors. Fortunately, this neglect of type 2 errors is changing. Recent studies have examined the statistical power and type 2 error rates in the published behavioral science, rehabilitation, and medical literature.[13-19,23-25] These investigations have consistently showed low power and a high probability of type 2 errors in the published literature. For example, Ottenbacher and Barrett[18] examined 100 data based investigations published in four major rehabilitation journals. Effect size values were labeled as large, medium, and small based on Cohen's[21] criteria. Data analysis showed that the median power to detect small, medium and large effect size values as statistically significant were 0.08, 0.26, and 0.56 respectively. These findings suggest that the probability of type 2 errors in the rehabilitation literature ranged from 44% for large treatment effects, to 92% for small treatment effects. In contrast, the probability of a type 1 error for any given statistical comparison was 5% ($p < .05$) or smaller.

The examination of statistical power in clinical and applied research has consistently showed low statistical power and the probability of high type 2 error rates.[26] The impact of low statistical power on the ability to develop a knowledge base to guide clinical practice are profound. As Lipsey[27] has noted, "statistical comparisons that lack power make for treatment effectiveness research that cannot accomplish its central purpose, ie, to determine the effects of treatment."

The problems of low power and statistical conclusion invalidity extend beyond the interpretation of individual research studies. Statistical conclusion invalidity can also affect the ability of a field to establish a useful and dynamic body of knowledge. The term "body of knowledge" is widely used in the research literature. The existence of a body of knowledge implies a series of related investigations examining a similar research question. Ideally, the results of accumulated clinical trials provide consensus on the research question of interest. The development of empirical consensus is an essential aspect of the research process, regardless of the method used to conduct individual investigations. The noted philosopher of science, John Ziman, has stated that the "goal of science is a consensus of rational opinion over the widest possible field."[28] This consensus is achieved through the replication of previous research. Ziman notes that "the results of repetitions of the same experiment are fundamental to the creation of any body of scientific knowledge."[28]

REPLICATIONS IN CLINICAL RESEARCH

Successful replication of quantitative research usually means that a null hypothesis that has been rejected in the original experiment will be rejected in a second investigation. This model of replication, based on the results of traditional statistical significance testing has been discussed in detail by Rosenthal[22] and is presented in table 2. The logic associated with the model in table 2 has recently been described by Goodman.[29] He states that "in practice, if a p-value is low enough to be considered significant, the observed effect is often claimed to be real, with the subsequent thought that a replication of the experiment should, with high probability, produce a similar statistical verdict.

The assumption made in using this model is that if a real treatment effect exists in the population, we should expect to produce statistically significant results (reject the null hypothesis) in the initial clinical trial and again when the investigation is repeated. If the null hypothesis is rejected in follow-up trials, we conclude that a successful replication has been achieved and that the theory tested, or intervention

Table 2: Standard Model of Replication Based on Statistical Significance Testing

	Original Study	
	$p > .05$	$p < .05$
Second study		
$p < .05$	Failure to replicate	Successful replication
$p > .05$	Failure to find effect	Failure to replicate

Adapted with permission.[22]