

and improve empirical consensus by implementing the following strategies. First, investigators should focus their interpretation on the data and the proper interpretation of statistical significance tests. The distinction between clinical and statistical significance is conveyed by reporting the results of significance tests as statistically significant rather than as significant. One repeatedly finds statements in the rehabilitation research literature that contribute to the confusion by failing to carefully label the results of statistical significance testing. For example, researchers often report only significance levels (ie,  $p < .05$  or  $p < .01$ ) without identifying the specific statistical test used or the obtained statistical value. This practice suggests that the researcher views statistical significance testing as an end, and not as a tool to be used in evaluating hypotheses. The ultimate judgment on the clinical significance of a research finding cannot be a statistical one.<sup>32,36</sup> Judgments on the clinical importance of a finding can be improved if the researcher reports summary data including, means, standard deviations, and frequencies and de-emphasizes the importance of statistical significance testing.

Second, researchers should supplement the results of any statistical significance test with the use of effect size values or confidence intervals. Cohen<sup>21</sup> has popularized the use of effect size measures appropriate for use with a wide variety of research designs and statistical tests encountered in rehabilitation research. The importance of effect size measures in clinical research is becoming widely recognized. In discussing research in education and psychology, Carver<sup>37</sup> recently observed that "nothing is more important in education and psychological research than making sure that the effect size of results is evaluated when tests of statistical significance are used." Along similar lines, the editor of the *Journal of Gerontology*<sup>38</sup> stated that "we are increasingly insistent that authors address the issue of effect magnitude, in both terms of the statistics reported and the discussion of the implications of the research. Any effect can be shown to be significant given sufficiently large sample sizes. The real question is whether or not the effect is important; measures of effect size help the researcher answer this question." Measures of effect size or confidence intervals should be routinely included in quantitative clinical trials reported in rehabilitation research.

Third, whenever the results of a hypothesis test are reported as statistically nonsignificant the researcher should report the statistical power of the test and the corresponding type 2 error probability. The probability of a type 1 error (usually  $p < .05$ ) is always known and is essential in interpreting the results of statistical tests when the null hypothesis is rejected, ie, the results are statistically significant. When the results are not statistically significant, researchers rarely provide any information on the probability that they made the correct decision. Lipsey<sup>27</sup> has argued that the failure to report information on type 2 errors and statistical power "degrades our ability to learn from research; to differentiate successful treatments from unsuccessful ones and find the keys to making the successful ones work even better." Statistical power values should always be reported for statistically nonsignificant results.

Fourth, statistical power should be an important consideration in planning clinical investigations. Recommendation three refers to the posthoc use of power to interpret statistically nonsignificant results. Statistical power can also be used to help ensure an adequate sample size is included in any clinical investigation involving hypothesis testing. Researchers who identify the significance level, power level, and estimated effect size, can use power tables to determine the sample size necessary to conduct a sensitive clinical trial with a high probability of correctly rejecting the null hypothesis.

Fifth, to develop a useful body of knowledge on rehabilitation interventions will require the replication of research findings and the synthesis of replicated results. Carver<sup>37</sup> argued that the best way to reduce the confusion associated with statistical significance testing is to replicate the findings. If chance (sampling error) is a possible threat to the validity of a research finding then the most effective way to reduce sampling error is to find the same result in a replication study. The importance of replication to the process of establishing scientific truth was illustrated in the widely publicized cold fusion experiment conducted at the University of Utah. Two investigators (Stanley Pons and Martin Fleischmann) reported the results of an experiment conducted at room temperature using an electrolytic cell. The investigators reported an excess of energy generated by cold fusion. Researchers around the world were unable to replicate the Pons and Fleischmann experiment. Scientific consensus could not be achieved on cold fusion and the result has been relegated to an interesting historic footnote.<sup>39</sup>

Replication research should be promoted by professional organizations and funding agencies. The effectiveness of various rehabilitation interventions will only be convincingly established through the development of empirical consensus. This consensus must be based on appropriately designed and accurately interpreted replicated clinical trials.

The field of medical rehabilitation cannot afford to misinterpret the results of quantitative clinical research at a time when theory development and treatment effectiveness are high priorities. Researchers and clinicians need to be educated in the proper application and interpretation of both quantitative and qualitative methods. Misinferences associated with statistical conclusion invalidity must be recognized and eliminated from our clinical research literature. De-emphasizing statistical significance testing will be an important step in re-establishing the usefulness and importance of quantitative methods in rehabilitation research.

The statistician, A. Wald, once observed that researchers often use statistics like a drunken man uses a lamppost, more for support than illumination. By developing a better understanding of the procedures used in making quantitative inferences we may lose some support, but we will see our facts more clearly because of the improved illumination.

**Acknowledgment:** The ideas in this article were greatly improved by discussion with Dr. Tom Matyas, Department of Health Behavioral Sciences, La Trobe University, Australia.

#### References

1. Deitz JC. Research: A systematic process for answering questions. In: Hopkins H, Smith HP, editors. Willard & Spackman's Occupational Therapy, 8th ed. Philadelphia: Lippincott, 1993:401-15