

# Marketing Papers

---

University of Pennsylvania

Year 2007

---

## Statistical Significance Tests are Unnecessary Even When Properly Done and Properly Interpreted: Reply to Commentaries

J. Scott Armstrong  
University of Pennsylvania, [armstrong@wharton.upenn.edu](mailto:armstrong@wharton.upenn.edu)

Postprint version. Published in *International Journal of Forecasting*, Volume 23, Issue 2, April 2007, 335-336.

Publisher URL: <http://dx.doi.org/10.1016/j.ijforecast.2007.01.010>

This paper is posted at ScholarlyCommons.

[http://repository.upenn.edu/marketing\\_papers/128](http://repository.upenn.edu/marketing_papers/128)

## **Statistical Significance Tests are Unnecessary Even When Properly Done and Properly Interpreted: Reply to Commentaries**

J. Scott Armstrong

This paper was published in the *International Journal of Forecasting*, 23 (2007), 335-336.

The three commentators on my paper agree that statistical tests are often improperly used by researchers and that even when properly used, readers misinterpret them. These points have been well established by empirical studies. However, two of the commentators do not agree with my major point that significance tests are unnecessary even when properly used and interpreted.

I am pleased that Paul Goodwin addressed the issue, "What is new?" Like Goodwin, I believe that the major point of my article is new to the vast majority of researchers and practitioners. Few forecasting researchers or practitioners are aware that there is no empirical evidence supporting the use of statistical significance tests. Despite repeated calls for evidence, no one has shown that the applications of tests of statistical significance improve decision-making or advance scientific knowledge. Schmidt (1996) offered this challenge: "Can you articulate even one legitimate contribution that significance testing has made (or makes) to the research enterprise (i.e., any way in which it contributes to the development of cumulative scientific knowledge)?" Schmidt and Hunter (1997) reported that no such cases have been reported and they repeated Schmidt's challenge.

In their commentaries, Herman Stekler and Keith Ord claim that statistical tests are needed for meta-analyses. Hunter and Schmidt (1996) demonstrate that this is false. They use the following example, which I have paraphrased: Assume that there is a true correlation of .3 between a measure of family composition and juvenile delinquency. What would happen if 50 studies were conducted to test for this relationship and the power of each test was .5 (a typical value in the social sciences)? In this case, half of the studies would conclude that there was a significant relationship at an alpha level of .05, and the other half would conclude that there was no relationship. Thus, by using tests of statistical significance, one would falsely conclude that there was no relationship. Hunter and Schmidt showed that prior to the development of meta-analysis, social science research was plagued with such faulty analyses, and that this hampered scientific progress. They demonstrated the superiority of meta-analyses that use effect sizes and confidence intervals. Becker (1987) contrasts the effects of using combined significance levels in three meta-analyses; She concluded that, when making inferences, effect-size analysis is superior to using

combined significance levels. Schmidt (1996) concluded that the use of meta-analysis will eventually lead researchers to abandon significance tests. This will bring scientific procedures in the social sciences more in line with those in the physical sciences.

Claims that significance tests are needed for meta-analyses were also refuted by the experience of those involved in summarizing cumulative knowledge for the *Principles of Forecasting* book (Armstrong 2001). None of the authors said that they needed significance tests.

Herman Stekler claims that I ignored papers expressing views that were contrary to mine. Schmidt and Hunter (1997) cite a number of such papers. However, I do believe that the proper way to address this question is to marshal the evidence rather than to take a vote among experts. I actively sought evidence that was contrary to my position.

Keith Ord's view is consistent with what I now believe to be an outdated principle from Armstrong (2001, p. 717):

**“13.29 Use statistical significance only to compare the accuracy of *reasonable* methods.**

Little is learned by rejecting an unreasonable null hypothesis. . . . [Statistical significance] can be useful, however, in making comparisons of reasonable methods when one has only a small sample of forecasts.”

No one has publicly challenged this principle, nor offered any evidence. However, after I conducted a further review of the evidence on statistical significance tests, I concluded that principle 13.29 needed to be revised to state instead that tests of statistical significance should not be used. One should assess effect sizes, and use confidence intervals and replications to assess confidence.

Koning, Franses, Hibon & Stekler (2005), violated principle 13.29 in that they tested an unreasonable null hypothesis (this may be true of many papers published in the *International Journal of Forecasting*). However, even if their analysis had been consistent with this principle, it would not have been helpful. And certainly it violates my revised version of the principle.

The social sciences have been led astray by significance tests. Scientific research can live without significance testing. Gigerenzer (2000, p 296) wrote, “Several years ago, I spent a day and a night in a library reading through issues of the *Journal of Experimental Psychology* from the 1920s and 1930s. This was

professionally a most depressing experience, but not because these articles were methodologically mediocre. On the contrary, many of them make today's research pale in comparison with their diversity of methods and statistics."

Some people think there is nothing new under the sun, while others marvel at new insights. The evidence on significance testing provided new insights to me. If indeed there were nothing new about significance tests, Koning et. al (2005) would not have been published. It was new to me that significance tests are harming the development of knowledge. In addition, significance tests take up space in journals. Finally, they violate what I believe to be one of the primary functions of statistics: to aid communication.

### References

Armstrong, J. S. (2001). *Principles of Forecasting*. Boston: Kluwer.

Becker, B. J. (1987), "Applying tests of combined significance in meta-analysis," *Psychological Bulletin*, 102, 164-171.

Gigerenzer, G. (2000), *Adaptive Thinking: Rationality in the Real World*. Oxford: Oxford University Press.

Hunter, J.E. & Schmidt, J. L. (1996) Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law*, 2, 324-347.

Koning, A. J., Franses, P.H., Hibon, M. & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, 21, 397-409.

Schmidt, F.L. (1996) Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115-129.

Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data, in Harlow, Lisa L., Mulaik, S. A. & Steiger, J. H. *What if there were no Significance Tests?* London: Lawrence Erlbaum.