Replications of forecasting research[☆]

Heiner Evanschitzky[a,*], J. Scott Armstrong[b]

[a] *Department of Marketing, University of Strathclyde, Glasgow, G4 0RQ, United Kingdom*
[b] *The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA*

Abstract

We have examined the frequency of replications published in the two leading forecasting journals, the *International Journal of Forecasting (IJF)* and the *Journal of Forecasting (JoF)*. Replications in the *IJF* and *JoF* between 1996 and 2008 comprised 8.4% of the empirical papers. Various other areas of management science have values ranging from 2.2% in the *Journal of Marketing Research* to 18.1% in the *American Economic Review*. We also found that 35.3% of the replications in forecasting journals provided full support for the findings of the initial study, 45.1% provided partial support, and 19.6% provided no support. Given the importance of replications, we recommend various steps to encourage replications, such as requiring a full disclosure of the methods and data used for all published papers, and inviting researchers to replicate specific important papers.
ⓒ 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Keywords: Replication research; Research policy; Census study

1. Introduction

Gardner and Diaz-Saiz (2008) replicated and extended research performed by Fildes, Hibon, Makridakis, and Meade (1998), which was itself an extension of the M-Competition study (Makridakis et al., 1982). Using changes in the estimation procedures, they concluded that the primary conclusion was supported but disagreed with one of the secondary conclusions. This demonstrates the value of replications, both in showing where we can gain confidence and in indicating areas which are in need of further research.

Experts have claimed that replication is vital to scientific progress (e.g. Hunter, 2001). Replications help to ensure that findings can be reproduced. Extensions go beyond that, and examine whether the findings can be generalized.

However, despite these benefits, relatively few of the papers published in various areas of management science are replications (Hubbard & Vetter, 1996; Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007). A number of reasons for this have been suggested. First, and perhaps most important, many studies in management science are unimportant, and thus it would be senseless to replicate them. Second, authors

seldom provide sufficient detail in the paper (or in response to requests) to allow for replication. And, third, reviewers seem to be biased against replications, either because they think that they do not offer anything new or because the results are not statistically significant.

The misinterpretation of null hypothesis testing procedures may also have undermined the perceived need for replication. Oakes (1986) showed that 42 out of 70 (60%) experienced academic psychologists falsely believed that an experimental outcome that is significant at the 0.01 level has a 0.99 probability of being statistically significant if the study were replicated. This is not true, and shows the low level of knowledge among some academics, which is possibly leading to the erroneous conclusion that replications are not needed.

Based on these observations, we examine the state of replication research in forecasting and then suggest ways to make further improvements with respect to replication.

2. The record of replications in the leading forecasting journals

The definitions of the central terms in this study are extensions of those employed by Hubbard and Armstrong (1994, p. 236). A replication is defined as "a duplication of a previously published empirical study that is concerned with assessing whether similar findings can be obtained upon repeating the study". Likewise, a replication with an extension is "a duplication of a previously published empirical research project that serves to investigate the ability to generalize earlier research findings". The latter goal of determining the range of conditions under which the findings do and do not hold up can also be addressed by studies conducting a series of experiments which are reported within the same article. However, such works lack independent assessment. Therefore, following Hubbard and Armstrong (1994), we only consider replications that were published as independent papers.

In forecasting journals, we find the following three types of replications: "real replications", "model comparisons", and "data re-analyses". While "real replications" are defined along the lines of the Hubbard and Armstrong (1994, p. 236) definition of replications, "model comparisons" are an application of a

previously published statistical analysis that is concerned with assessing whether a superior goodness-of-fit can be obtained, comparing the original statistical model with at least one other statistical model. "Data re-analyses" the third category of replications in forecasting journals, can be defined as an application of previously published data that is concerned with assessing whether similar findings can be obtained using a different methodology with the same data or a subsample of the data. For example, McCullough (2000) showed that the selection of a software program for a given method had an important effect on the results. It is worth noting that the aforementioned three types of replications in forecasting are not mutually exclusive, and are frequently used in conjunction.

We did a frequency count of the three types of replications appearing in the *International Journal of Forecasting (IJF)* and the *Journal of Forecasting (JoF)*. The numbers are estimated based on a census of all 929 empirical articles published in the two journals over the period 1996 through 2008. Two independent raters classified each of these works to determine the publication incidence of replications and extensions. The raters were in agreement for 79% of their initial classifications,[1] which is slightly less than the 88% agreement rate reported by Evanschitzky et al. (2007) and the 90% rate reported by Hubbard and Armstrong (1994).

The results of the count show that the most prominent type of "replication" seems to be model comparisons (26.9% of replications on average for *JoF* and *IJF*), followed by data re-analyses (17.6%). The average of "real" replications was 8.4% for the two forecasting journals. There were no clear trends over the 13-year assessment period.

Comparing the number of replications and extensions in the forecasting journals with those being published in other areas of management science, we see similar results. Hubbard and Vetter's (1996) analysis of replications of empirical studies showed that the average replication rate varied between 2.2% of the empirical papers (*Journal of Marketing Research*) and

---

[1] The agreement is for all three categories. It is worth noting that we did not find any pattern of disagreement between the independent raters. When a disagreement was found, we first asked the raters to discuss it amongst themselves. Then, based on their suggestion, one of the authors of this paper made the final decision on how to rate a particular study.