

# Time-Series Analysis: Assessing the Effects of Multiple Educational Interventions in a Small-Enrollment Course

Aaron R. Warren

*Department of Mathematics, Statistics, & Physics  
Purdue University North Central  
1401 S. US-421 Westville, IN 46391  
awarren@pnc.edu*

**Abstract.** Time-series designs are an alternative to pretest-posttest methods that are able to identify and measure the impacts of multiple educational interventions, even for small student populations. Here, we use an instrument employing standard multiple-choice conceptual questions to collect data from students at regular intervals. The questions are modified by asking students to distribute 100 Confidence Points among the options in order to indicate the perceived likelihood of each answer option being the correct one. Tracking the class-averaged ratings for each option produces a set of time-series. ARIMA (autoregressive integrated moving average) analysis is then used to test for, and measure, changes in each series. In particular, it is possible to discern which educational interventions produce significant changes in class performance. Cluster analysis can also identify groups of students whose ratings evolve in similar ways. A brief overview of our methods and an example are presented.

**Keywords:** time-series analysis, cluster analysis, research methodology, electromagnetism

**PACS:** 01.40.Fk, 02.50.Ey

## INTRODUCTION

Pretest-posttest designs are commonly used to assess the effects of educational interventions [e.g., 1,2]. Recently, a more sophisticated between-subjects design was employed by Sayre & Heckler to identify dynamical changes in student performance [3]. However, this design requires a large number of participants in order to obtain the adequately large quasirandom subsamples necessary for each measurement.

The need for quantitative assessment methods appropriate to small- $N$  situations is significant. Small enrollment courses have difficulty using conventional study designs due to sample size restrictions. Moreover, adaptations of instructional materials and methods designed for specific student populations cannot be properly assessed in such courses. An ability to quantify the effect of an instructor's unique array of educational interventions on students' learning, or determining how the effect on students' learning is modulated by intrinsic factors, would provide a powerful feedback mechanism for the improvement of instruction in such courses. A supplementing of qualitative techniques with quantitative research methodologies for small- $N$

courses would therefore be a boon to course assessment and education in general.

In this work we aim to address two questions. First, how can we identify and measure the effects of specific interventions (lectures, labs, homework assignments, etc.) on student performance in a small- $N$  environment? Second, how can we determine whether the effects of specific interventions vary according to some intrinsic factors such as students' gender or major?

## STUDY DESIGN & ANALYSIS

In this section, we describe a method for collecting data suitable for time-series analysis and outline some of the features of a particular type of time-series analysis known as *autoregressive integrated moving average* (ARIMA) modeling. We also provide an example of cluster analysis applied to time-series data.

### Data Collection

Data were collected from the Spring 2009 PHYS 221: General Physics II course at Purdue North Central. This is an algebra & trigonometry-based

course taught in the evenings covering electrodynamics, optics, and modern physics, with an enrollment of 38 students. At the end of every lecture and lab, each student completed and turned in a “Physics Journal” entry. The only class sessions when journal entries were not done were exam days, of which there were three during the semester.

In a journal entry, a student is asked to rate his/her confidence in each of the answer options for 8 multiple-choice questions. Students are given 100 Confidence Points for each question to distribute among the answer options. The confidence rating indicates the self-reported perceived likelihood that an answer option is correct. Ratings employ a scale of 0-100, with 100 indicating absolute confidence that an answer option is correct and 0 indicating absolute confidence that an answer option is incorrect. The set of 8 questions is the same for all journal entries, and included several items from the CSEM [4]. Over the semester, there were a maximum of 30 journal entries per student, and a maximum of 1140 journal entries in total for the course. Due to absences, the actual number of journal entries was 955, giving a response rate of 84%. In this paper, we will deal only with an example analysis of one item, CSEM #23. A full report on our study is in preparation.

Confidence ratings on each answer option were averaged over the entire class. There are several ways to average the individual ratings in order to account for absences. For example, since our goal is to determine the performance impact of each intervention, for the  $i^{\text{th}}$  journal entry by a student we may choose to include it in the class average only if the student was also in attendance for the  $(i-1)^{\text{st}}$  journal entry. This increases the degree of continuity in the responses by reducing the effect of absences, although it incurs a cost as the effective response rate drops to 73%.

An alternative approach is to simply include all journal entries in the class averages and treat the absences as an additional source of noise. The analyses presented in this study were done via both approaches, as well as a few other variations, and the results were essentially identical. Here, we will report results only for the latter approach.

This study design has a high risk of suffering practice effects. The between-subjects study by Sayre & Heckler was able to identify three features in the evolution of student responses to several CSEM items; peaks, decays, and interferences. As our analysis will show below, practice effects seem to prevent us from detecting decays, although we are still able to identify peaks and interferences. This limitation is an unavoidable consequence of working with a small-enrollment course.

## ARIMA Modeling

The class-averaged ratings for each answer option can be viewed as a ‘Dow Jones Index’ for the class, and in fact when graphed they do appear similar to financial time series. A natural thought is to try using ARIMA models, commonly employed in quantitative finance [5], ecology [6], and genetics [7], to model our data. In this section we briefly introduce both ARMA and ARIMA modeling. All analyses were performed using the *R* statistical analysis environment [8].

The purpose of any time-series model is to define an equation that can well-approximate the observations at each time,  $X_t$ . An ARMA( $p,q$ ) model attempts to fit an equation of the form:

$$X_t = \alpha_0 + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (1)$$

where  $\alpha_0$  is the mean,  $\varepsilon_t$  is a white noise series,  $p$  and  $q$  are positive integers,  $\varphi_i$  are the autoregressive (AR) coefficients to be estimated by the fitting, and  $\theta_i$  are the moving average (MA) coefficients to be estimated also. In general, at least 5 observations per parameter are required for the estimates to be at all reliable. An ARMA( $p,q$ ) model has  $p+q+1$  parameters.

ARMA model identification, estimation, and diagnostic checking are codified by the Box-Jenkins procedure [9], which we outline here. First, ARMA modeling is only appropriate for stationary time-series, which have a constant mean. When the mean of a time-series exhibits trending, it is made stationary by differencing. That is, instead of fitting an ARMA model to the series  $\{X_t\}$ , the series of differences  $\{Z_t\} = \{X_t - X_{t-1}\}$  is used. Depending on the type of trend (linear, quadratic, etc.), differencing may need to be applied multiple times to stabilize the mean. The fitting of an ARMA( $p,q$ ) model to a series that has been differenced  $d$  times is called an ARIMA( $p,d,q$ ) model.

Once suitably differenced, an attempt is made to identify the optimal orders  $p$  and  $q$  for the AR and MA processes. This is done by analyzing the autocorrelation functions (ACF) and partial autocorrelation functions (PACF), which measure repeating behavior within the time-series. These functions typically suggest a good initial model. Other models consistent with the various autocorrelation functions are also constructed. Our final choice of model is based on the *corrected Akaike information criterion* (AICC), which assesses the benefits of improving goodness of fit versus the costs of adding additional parameters.

Finally, a model undergoes diagnostics to determine its suitability. Ideally, a model should explain all systemic features of the series, with any

residuals exhibiting the behavior of Gaussian white noise. This is partially addressed by using the Ljung-Box test to determine whether the residuals are uncorrelated, which is a necessary condition to qualify as white noise.

Time-series may also have structural shifts, where the series changes its behavior due to some exogenous intervention. In this case, the series is broken up into piecewise sections, each of which must be modeled separately from the rest of the series. Possible structural shifts are often identified by eye. These identifications are tested by constructing a model for the section of the series preceding the proposed structural shift, using that model to forecast 95% confidence ranges for several points beyond the end of the section, and observing whether the actual data points lie within that range. If not, it is likely that a structural shift occurred. The next section of data is then modeled separately. The estimated effect of the shift is calculated by comparing the difference in values for the two models at the time of the shift.

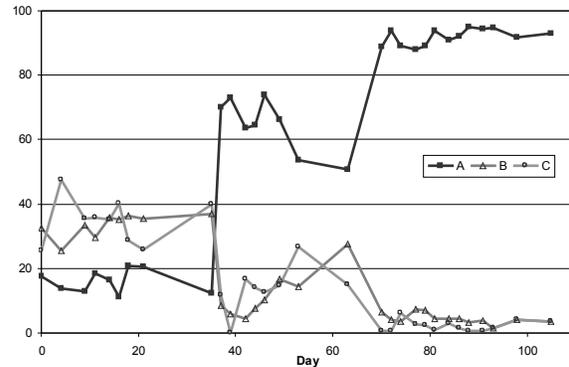
Our study employs the Box-Jenkins procedure to identify and quantify structural shifts. We argue that these shifts represent peaks and interferences as identified in the work of Sayre & Heckler. Below we present an example of this.

### Example Analysis

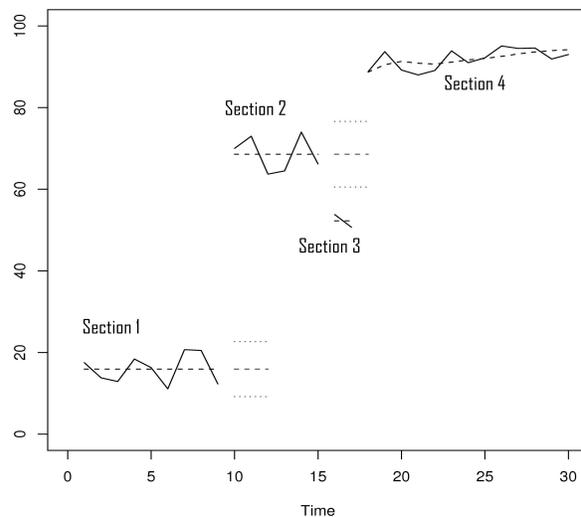
In Figure 1 we show the time-series of the class-averaged confidence rating for three answer options (*A*, *B*, and *C*) to CSEM #23. This question asks students to determine the direction of the net magnetic field at a point between two current-carrying wires. Two answer options (*D* and *E*) are not plotted as they were minimal throughout the study. We will model the series for the correct answer *A*.

Looking at series *A*, we identify possible structural shifts at Days 37 and 70, and possible interference at Days 53 and 63. Day 37 was the lecture covering the right-hand rule for current-carrying wires, and included an in-class group-work assignment on the topic. Day 70 was the first class after the mid-term exam covering magnetism, and therefore includes the effects of students' exam preparations. Other relevant dates were Day 39, when a lab involving testing experiments for the various right-hand rules was run, and Day 42 when a hybrid online & paper homework assignment on the topic was due.

These visual identifications are supported by the ARIMA modeling. The models and forecasts used to test for structural shifts are shown in Figure 2. Sections are labeled 1 (pre-lecture), 2 (between lecture and interference), 3 (interference), and 4 (post-exam).



**FIGURE 1.** Time-series of confidence ratings for answer options *A*, *B*, *C* from CSEM #23. The correct answer is *A*. Error bars not shown due to graphical limitations. Time is measured in Days from the beginning of the semester.



**FIGURE 2.** Series *A* with ARIMA models and forecasts included. The independent variable Time counts the number of class sessions elapsed. Observations are solid black lines, models & forecasts are dashed black lines, 95% confidence ranges on forecasts are dotted grey lines.

ARIMA models assume that observations are collected at regular intervals. Although our observations are not made at regular intervals in actual time, they are made at regular intervals in class-time. For the purposes of analysis, time is therefore measured by the class session number (i.e., class sessions 1, 2, 3, etc.) instead of days elapsed.

The series for Section 1 exhibits no trending. Analysis of the ACF and PACF for Section 1, and comparison with the AICC for other models, suggests an ARIMA(0,0,0) model:

$$X_t = \alpha_0 + \varepsilon_t \quad (2)$$

i.e., a constant mean with white noise. The mean is estimated to be  $\alpha_0 = 15.94 \pm 1.13$ . Diagnostics, including the Ljung-Box test, indicate this model provides an adequate fit. Forecasts for the next three

class sessions beyond this section (Days 35, 37, 42) are generated using this model. These forecasts and their 95% confidence ranges are included in Figure 2. As the data from these three class sessions lie well outside of this range, we infer that a structural shift occurred at the end of Section 1.

Section 2 is similarly fit with an ARIMA(0,0,0) model, where  $\alpha_0 = 68.57 \pm 1.64$ . As the observations for Days 53 and 63 lie well outside of the forecasted range based on this model, we infer that there is some significant source of interference during Section 3.

Section 3 is too short to model, having only 2 observations. In Figure 2 we have simply plotted the average,  $52.25 \pm 1.10$ . This interference occurred during the lectures covering inductors, transformers, and AC circuits. It is possible that learning the right-hand rule for induction caused the interference, shifting confidence toward options B and C. It is at this point that a qualitative study would be useful to determine the underlying cause of the interference and propose ways to minimize it.

Section 4 is fit with an ARIMA(0,1,1) model:

$$X_t - X_{t-1} = \alpha_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (3)$$

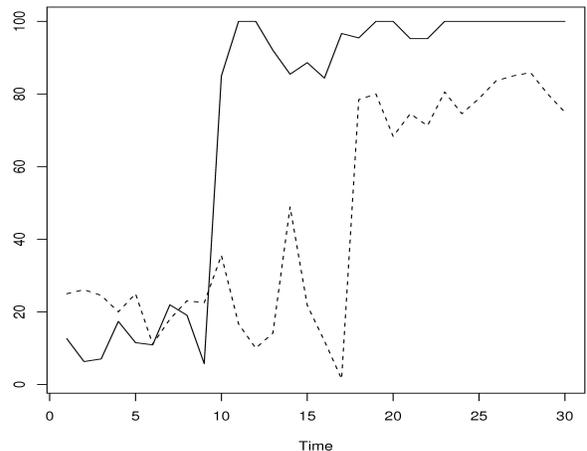
where  $\alpha_0 = 0.37 \pm 0.15$  is the drift in  $X_t$ , likely due to a practice effect, and  $\theta_1 = -1.00 \pm 0.25$ . This is a moving-average process, where prior noise fluctuations affect the future mean of the series. We believe this is due to absences, as a student who missed one class was unlikely to miss the next. Thus, the noise in class performance gains due to an absence was likely to be reversed at the next class session, explaining the estimated MA coefficient of -1.00.

Our analysis detects peaks and interferences as expected by the Rescorla-Wager model employed by Sayre & Heckler, but is unable to identify any decays due to practice effects in our design. One other limitation of our design is the potential for priming. After completing several journal entries, students may pay more attention to material that is related to the journal items. Thus, the performance gains due to the lecture may be artificially enhanced.

In summary, the ARIMA analysis indicates the lecture boosted performance by  $52.63 \pm 2.77$  points, the lab and homework assignments were ineffective, the lectures on AC circuits temporarily reduced performance by  $16.32 \pm 2.74$  points, and the exam increased performance by  $20.23 \pm 3.73$  points.

Finally, we perform a hierarchal cluster analysis of individual student ratings for option A. Two clusters identified at the 0.90-level include all subjects, with Cluster 1 containing 24 students and Cluster 2 containing 14 students. Average ratings for each cluster are shown in Figure 3. As can be seen, the clusters differ depending on whether the lecture or the exam was the primary mechanism for increasing

performance. Clusters are independent of gender ( $\chi^2 = 2.263$ ,  $df = 1$ ,  $p = .132$ ) and major ( $\chi^2 = 1.208$ ,  $df = 1$ ,  $p = .272$ ).



**FIGURE 3.** Average ratings for the two clusters identified by hierarchal cluster analysis of individual student responses to answer option A. Cluster 1 = solid line, Cluster 2 = dashed line.

## CONCLUSION

ARIMA time-series analysis provides a method for detecting and measuring the effects of multiple educational interventions, even in small-enrollment courses where standard study designs are impossible. Although performance decays are removed by practice effects, peaks and interferences can be observed. Moreover, cluster analysis can also be employed to test whether factors intrinsic to the students modulate the effects of specific interventions.

## REFERENCES

1. E. F. Redish, J. M. Saul, and R. N. Steinberg, *American Journal of Physics* **66**, 212-224 (1998).
2. R. R. Hake, *American Journal of Physics* **66**, 64-74 (1998).
3. E. C. Sayre, and A. F. Heckler, *Phys. Rev. ST Phys. Educ. Res.* **5**, 013101 (2009).
4. D. P. Maloney, T. L. O’Kuma, C. J. Hieggelke, and A. V. Heuvelen, *American Journal of Physics* **69**, S12-S23 (2001).
5. R. S. Tsay. 2005. *Analysis of Financial Time Series*. Hoboken, NJ: Wiley.
6. D. L. Druckenbrod, *Can. J. For. Res.* **35**, 868-876 (2005).
7. Z. Bar-Joseph, *Bioinformatics* **16**, 2493-2503 (2004).
8. R Development Core Team, *R: A language and environment for statistical computing*, <http://www.R-project.org> (2009).
9. G. E. P. Box, and G. M. Jenkins. 1976. *Time series analysis: Forecasting and control*. San Francisco: Holden Day.