Data Reduction Methods Applied to Understanding Complex Learning Hypotheses

Philip I. Pavlik Jr.
ppavlik@andrew.cmu.edu
Human Computer Interaction Institute, Carnegie Mellon University

Abstract. Modern learning science researchers are facing a flood of data as it becomes easier and easier to collect multiple streams of information from students before, during, and after learning experiments. While oftentimes these experiments do experimentally manipulate specific variables to improve responses on a posttest, these experiments are also interested in how the many related student factors explain who responds to the treatment and why. This poster introduces a recent experiment and explains how the data were analyzed using a combination of exploratory factor analysis (using SPSS) and exploratory structural equation modeling (using Tetrad) to partially refute a theoretical hypothesis and reveal a new explanation for further testing.

1 Introduction

A recent line of experimentation with the FaCT System [1] for vocabulary drill practice in Chinese is focusing on how metacognitive and motivational factors may be important in various aspects of student learning. Specifically, this line of research is trying to dig deeper into various reliable correlations that were uncovered in earlier work. In this earlier work students were ask for self reports of strategy use (e.g. forming verbal linkages rather than just repeating items). A consistent finding in 2 experiments was that students in the less difficult conditions (higher percent correct) reported using more of these verbal strategy links rather than using a more repetitive strategy [2]. My initial hypothesis was that this pattern of results suggested that higher percent correct (more review practice of the same flashcards) reduces the cognitive load on the user, thereby allowing the individual to better use learning strategies. Since strategy use for learning paired-associates has often been a successful way to improve learning [3], I wanted to test whether the process based explanation above was correct, since if it should be supported, it would indicate that any further ways we could decrease cognitive load might encourage further strategy use by students thereby leading to further learning gains.

The data we collected before the study included the full Motivated Strategies for Learning Questionnaire (MSLQ). During the vocabulary practice we gathered all data on recall and performance, and also surveyed students on a 5-point Likert scale to ask them to rate how difficult practice was recently, how useful practice was recently, and how much they were able to use strategies for learning recently. Further, during this middle portion they were randomized into either easier or more difficult practice conditions (more or less review). Following practice they responded to a 32-item questionnaire that specifically addressed their reactions to the practice sessions.

2 Analysis and Results

Analysis began with exploratory factor analysis on both questionnaires. Because I had a limited sample size (49 students) it was clear that we would not be able to meaningfully discover all the MSLQ factors found in prior research [4], which suggests that 15 subscales can be differentiated from the 81 items. I used principal components FA (SPSS factor analysis procedure) with oblimin rotation (which allows correlated factors to be found) to initially reduce the MSLQ data, and then maximum likelihood FA (which produces fit statistics) to find a best fitting model of the MSLQ factors that included mostly high communality items. Four factors were found. Data from the post-questionnaire was similarly reduced to 3 factors. The factor scores for the 4 MSLQ and 3 post-questionnaire along with all the other relevant data (performance data, Likert survey averages during practice, difficulty condition the student was in) were entered into the GES (greedy eliminative search) algorithm in Tetrad software program. This particular structural equation modeling (SEM) algorithm was ideal for my purposes because it provided a theoretically unmotivated way to search for a best way to reduce the patterns of the data into something understandable. Our hypothetical model (easiness leads to strategies) which required only a single edge from difficulty during learning to strategies during learning fit more poorly ($d.f. = 81$, $X^2 = 87.6$, $p = 0.298$) than an identical model with no requirements on this edge in the model ($d.f. = 80$, $X^2 = 81.83$, $p = 0.422$).

The SEM results in the poster will show the differences between the competing models and also illustrate the power of this approach for drawing hypotheses from large sets of poorly organized data or from data without strong preexisting theoretical relations. In particular this sort of analysis reveals one way contextual data about student data in repositories such as DataShop can be used to make conclusions about learning.

Acknowledgements

This research was supported by Ronald Zdrojkowski and was also made possible with the assistance and funding of the Pittsburgh Science of Learning Center (NSF-SBE) #0836012.

References

[1] Pavlik Jr., P.I., Presson, N., Hora, D.: Using the FaCT System (Fact and Concept Training System) for Classroom and Laboratory Experiments. *Inter-Science Of Learning Center Conference,* 2008. Pittsburgh, PA,

[2] Pavlik Jr., P.I.: The microeconomics of learning: Optimizing paired-associate memory. *Dissertation Abstracts International: Section B: The Sciences and Engineering,* 2005, 66(10-B)**,** p. 5704.

[3] Atkinson, R.C., Raugh, M.R.: An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning & Memory*, 1975, 1(2)**,** p. 126-133.

[4] Artino, A.R., Jr.: Review of the Motivated Strategies for Learning Questionnaire. 2005.