(2007). Getting things in order: Collecting and analysing data on learning. In F. E. Ritter, J. Nerb, E. Lehtinen, & T. M. O'Shea (Eds.), In order to learn: How the sequences of topics affect learning. 81-92. OUP.

Getting things in order: Collecting and analyzing data on learning

Frank E. Ritter Josef Nerb Erno Lehtinen

To appear in: Ritter, F., Nerb, J., O'Shea, T., & Lehtinen, E. (Eds.). (in preparation). *In order to learn: How ordering effects in machine learning illuminates human learning and vice versa*. New York, NY: Oxford University Press.

Frank Ritter +1 814 865-4453 frank.ritter@psu.edu

Josef Nerb +49 761 682-376 nerb@ph-freiburg.de

Erno Lehtinen +358-2-3338824 -8830(FAX) erno.lehtinen@utu.fi

Abstract

Where shall we start to study order effects in learning? A natural place is to observe learners. We present here a review of the types of data collection and analysis methodologies that have been used to study order effects in learning. The most detailed measurements, such as simple reaction times for completing a task, were developed and are typically used in experimental psychology. They can also form the basis for higher level measurements, such as scores in games. Sequential data, while less used, are important because they retain the sequential nature of observations, and order effects are based on sequences. These records can include eye movements, subjects' spoken-aloud thoughts as they solve problems (verbal protocols), and records of task actions. In areas where experimental data cannot always be obtained, other observational techniques are employed such as surveys. Once gathered, these data can be compared with or "cooked down" into theories, which can be grouped into two types: (a) Static descriptions that describe the data without being able to reproduce the behavior, examples includes simple behavior grammars and Markov model. (b) Process models that perform the task that subjects do and thus make predictions of their actions. These process models are typically implemented as a computational system. They provide a more powerful, dynamic description, but one that is inherently more difficult to use.

Acknowledgements

Georg Jahn, Mike Schoelles, and William Stevenson provided useful comments on this chapter, Mike particularly the Appendix.

4.1 INTRODUCTION

Where shall we start to study order effects in learning? A natural place is with data. We review in this chapter several of the types of data for studying order effects in learning, and a selection of existing, well-established methodologies for collecting and studying such data. Some of these methodologies themselves are often underused, however, so this chapter may encourage the use of these deserving (but often expensive in time or equipment) data collection and analysis methodologies. We present approaches from psychology, education, and machine learning, which—as we believe—can be fruitfully applied in other disciplines.

We are interested in data that can show that order effects occur and give us insight into how they occur. In addition, of course, we would also like the data to be robust, that is, the data should be reproducible and reliable. This will sometimes imply special techniques for data gathering.

We will see several themes and issues in exploring the types of data that can be used. First, there is a need to keep the sequential nature of the data intact to study sequential phenomena. Second, there is a trade-off between the detail of the data and the amount of data that can be gathered and analyzed with a given amount of resources. For example, you can see that chapters here that gather a lot of data per subject and do very detailed analyses use fewer subjects than studies that gather less data per subject or perform more automatic analyses. Third, we will present several data types and a discussion of corresponding, appropriate analysis techniques. Fourth, we turn to the issue of different experimental designs for studying order effects. The end of the chapter discusses how your data can be embedded within broader theories of human learning and problem solving.

4.1.1 Retaining the sequential nature of the data

It is not strictly necessary to keep the sequential order of the data to study order effects themselves. Order effects can often be found simply by looking at how subjects perform after receiving stimuli in two different orders. It is necessary to keep the sequential aspects of the data in mind to be able to observe where and when these order effects appear (they might be practically very important as well!). In addition, and theoretically more important, understanding how the order effects occur, is greatly assisted by having intermediate measures of performance that retain the sequential nature of behavior.

Figure 1 gives an illustration of one of several possible order effects. It shows how performance (typically an inverse of response time) might vary with two different learning orders. If you measure after two units, there is not an order effect because the stimuli are not equivalent. If you measure after three or four units of time, there is an order effect. At five

units of time, there is not an order effect for E, but there remain the difference performance effects on the intermediate stimuli (D is most prominent), and there are likely to be residual effects in many learning systems.

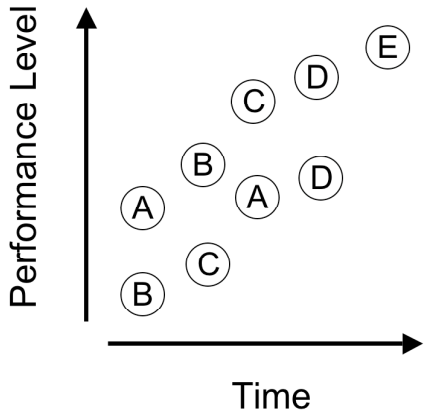

Figure 1. Order effects are visible after measuring after ABC vs. BCA and after ABCD vs. BCAD, but there is no effect after ABCDE vs. BCADE.

Retaining the sequential nature of data is not dependent upon what kind of data are gathered, although most types of data have traditionally either discarded the sequential information (e.g., reaction times), or traditionally retained the sequential order of the data (e.g., verbal protocols). In the case presented in Figure 1, the data needs to be retained for the units as well as their order. To be sure, you always can collect sequences of elementary data, such as sequences of reaction times, of test scores, of verbal utterances, and so on, and keep them as sequences. We will present examples of those data sequences later.

Recently there have been steps to extend the use of sequential data. Exploratory Sequential Data Analysis (ESDA) in human-computer interaction studies (Sanderson & Fisher, 1994), and in the social sciences in general (Clarke & Crossland, 1985) allows you to see intermediate order effects.

4.1.2 Data granularity

Of course, choosing the appropriate level of data to examine is crucial. If you use detailed enough data, you can often see a large amount of intermediate order effects, as you see learners on different paths come to the same performance (see again Figure 1). VanLehn's results (this book) suggest this is possible. Finer grained data will also provide more insight into the learning mechanisms.

There are trade-offs, however. More data often means that data collection will get more cumbersome and that the analysis becomes more complicated. In addition, as we know from

the statistical theory of mental test scores (Lord & Novick, 1968), single observations are less reliable than an aggregation over a set of similar observations. Thus, using aggregated data by collapsing blocks of multiple observations over time increases the statistical power of your research at the cost of ignoring potential interesting interactions within the collapsed blocks such as order effects. It was often said by Newell and Simon (personal communication) that the most interesting trials were the practice trials before starting the experiment proper, because these were where subjects[1] learned.

4.2 TYPES OF DATA AND THEIR GATHERING AND ANALYSIS

We will examine several types of data in detail. This is not to say that there are not other types, just that these are either the most natural or are particularly good examples. This will include simple quantitative measurements, qualitative measures, measures from students, and data from models and automatic learners.

4.2.1 Simple quantitative measures

Measures such as time to complete a task (response times) and quality of performance (percent correct) are not the most exciting way to study order effects, but they are a good place to start because they are simple and clear. When they are taken at the end of two stimuli orders they can provide the first indication that order effects are occurring in learning. Learning curves, such as shown in Figure 1, are often generated from repeated assessing of those simple performance measures. Reaction times are also among the most traditional ways of studying behavior. Especially in applied research, time to perform a task can be crucial because it represents money or consumes other resources.

Part-task training is a domain where time to learn and performance are the measures typically examined. Here, complex tasks are decomposed into smaller units that can be efficiently trained in isolation. The goal, then, is to find a decomposition and an optimal training sequence for those smaller units that minimize the cost of learning the total task (see e.g., Donchin, 1989 for a relatively complex task; and Pavlik, this volume, for a relatively simple task example that examines only training order, not decomposition).

Other often used simple measures include counting of correct and incorrect responses. Many of the chapters here start with these measures. In general, these simple quantitative measure

[1] We have followed Roediger's (2004) use of "subjects" to refer to subjects because experimenters are also participants.

can be a useful indicator and summary of order effects that you will often wish to see, but they will not tell you much about how and why the effects occurred.

4.2.2 Derived measures

Simple measures can be combined to create more complex measures. A good example of a derived measure is velocity (as it is derived from distance and time). Examples in behavior would include sums, differences, and ratios of reaction times or such manipulations of other kinds of indirect measures. We can note several interesting kinds of derived measures to keep in mind, which we explain next.

Hybrid measures

Combining several measures (e.g., scoring 5 points for each second to complete a task and 10 points per widget) are often used to create scores provided to people learning a procedural task. The highly motivating learning experiences called video games, for example, often use them. A problem with these measures is that they are ad hoc, and thus they often fail to meet the assumptions necessary for inferential statistical tests (for an account when and how several measures can be combined meaningfully see Krantz, Luce, Suppes & Tversky, 1971; or other good statistics books). These scores are nevertheless common practice in the classroom, for example, many tests give points for knowing different types of knowledge. From an applied point of view, they may be initially useful as a summary of performance. For further theoretical analysis, however, you will need to keep the components separate and check for possible interaction between the parts before you build summary scores.

Change as a measurement

In order to change the impact of learning on performance, it is sometimes useful to compute differences in performance. This can be differences in time to successfully complete a task (has learning changed the speed of performance?), differences in error rates, or differences in other quantitative measures you are using. Turn taking is another derived measure, for a turn is defined in relation to another action. But be always aware of problems in inferential statistics using differences as a dependent variable!

Other interesting change measures include interaction patterns. These can be represented with a variety of grammars (e.g., Olson, Herbsleb, & Rueter, 1994), and can be analyzed to find precursors for behaviors using lag sequential analyses (e.g., Gottman & Roy, 1990)

4.2.3. Applying codes to measures: Qualitative Measures

In order to study the impact of learning, sometimes it is useful to study how performance changes qualitatively, such as strategy shifts. This can be done by building meaningful categories and coding the subject's behavior. These codes (categories) can then be analyzed as other data.

An example of this type of study is research dealing with the level of aspiration in a series of tasks with varying difficulty. In a study by Salonen and Louhenkilpi (1989) students solved anagram tasks in an experimental situation where they had to select a series of tasks from five levels of difficulty. Students had restricted time for each task, and they had to select and solve several tasks. In the middle of the series students were given superficially similar but impossible tasks. The effect of induced failures was different for students with different motivational tendencies. Some students slightly lowered their aspiration level after the failures but raised it again after the later success. Other students responded to failures by decreasing their aspiration level and kept selecting the easiest tasks independently of occasional success during later trials. Students' selections were videotaped and they were interviewed after each selection. This qualitative data were combined with the quantitative data of selecting sequences and successes. (This data and analysis approach is similar to work reported in VanLehn's chapter.)

In another study (Lehtinen, Olkinuora & Salonen 1986) students solved problems of addition and subtraction of fractions. In this study there were also impossible tasks in the middle of the series. Qualitative differences of the problem solving processes before and after induced failures were observed. Some students solved the problems without showing any effect of the induced failures, whereas other students became worse in their problem solving processes after the induced failures. This might suggest possible mechanisms for order effects in learning (in this case, emotional responses, also see Belavkin & Ritter, 2004), and highlights the effect of order on motivation and the role of motivation in learning.

4.2.4 Protocols and theoretical frameworks

All measurements are taken within a theoretical framework, even if one might not be aware of it. Some measurements, however, are taken within a larger and more explicit framework than others. Protocol data, sequences of behavior, typically provide a rich account of all kind of behavioral observations. Protocols are an important area of measurement that can be used to study learning that often need to have their measurement theory made more explicit.

Protocols allow us to look at the time course of learning and are usually able to provide additional information on processing and learning, which many types of data do not address. Many types of protocols are related to an explicit theoretical framework and form an

important method for studying learning processes. Examples of protocol data include sequential recording of verbal utterances during problem solving (e.g., VanLehn this volume), mouse and keyboard events whilst working with a computer (e.g., Pavlik, this volume,; Swaak & De Jong this volume; Scheiter & Gerjets, this volume), or eye movements during reading. To find regularities within such vast records you need a theoretical framework to provide guidance.

Each type of protocol data comes with a theoretical framework of how and why they can be used. Verbal protocols—often called talk-aloud-protocols—are perhaps the best known. Verbal protocols are taken within a strong framework (Ericsson & Simon, 1993). They make several explicit assumptions about how subjects can access working memory, and how they can report through "talking aloud." Verbal protocols can provide cues about what information subjects are using, and point to strategies that were employed by subjects. Eye movements have been studied as well (from early work summarized by Monty & Senders, 1976, Rayner, 1989, to more recent work such as Byrne, 2001, Anderson, Bothell, & Douglass, 2004, and Hornof & Halverson, 2003), and help us understand how order effects occur by suggesting what information subjects have paid attention and in what order. These protocols can include mouse movements where they are different than task actions, but these, too, require a theory to support a less direct measurement theory (Baccino & Kennedy, 1995; Ritter & Larkin, 1994).

In educational psychology the units of analyses have typically been larger than in experimental cognitive psychology, and thus the data acquisition methods are somewhat different. They can include stimulated recall interviews where students, for example, watch a videotape of the sequence of their own activities and try to explain the meaning and intention of different acts (Järvelä, 1996). (This is a type of retrospective verbal protocol, Ericsson & Simon, 1993).

So far, gathering and analyzing all types of protocols have been difficult enough that they have not been used as often as one might like. However, the theories supporting the use of protocols are robust and protocols can detail the micro structure of how order effects could occur and often provide insight into the mechanisms that give rise to order effects.

4.2.5 Machine learning data

The behavior of machine learning algorithms, for example, as noted by Cornuéjols (this volume), can be examined in pretty much the same way as human subjects (Cohen, 1995; Kibler & Langley, 1988). Very often the same measures can be taken. Machine learning algorithms, however, are nearly always easier to study than human subjects because the learning algorithms are typically faster to run than subjects, and they do not have to be