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.

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.

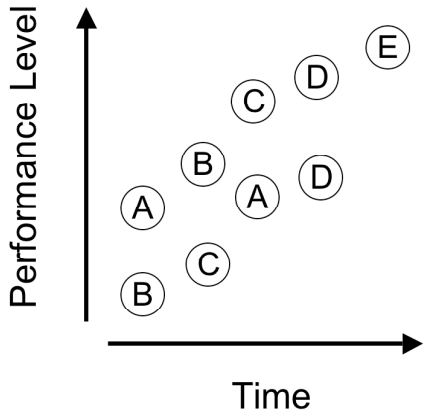

Figure 1. Order effects are visible after measuring after ABC vs. BCA and after ABCD vs. BCAD, but there is no effect after ABCDE vs. BCADE.

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

recruited to do the task. You can easily control for confounding variables and you do not have to be concerned with factors related to the social psychology of the experiment (such as demand characteristics or experimenter expectancy effects). In addition, it is easy to reset the learning algorithm and run it over a new input stimuli set or with changes to the model's parameters (representing different subjects or subject populations). The analyst can also take additional measurements of the internal state of the learner directly, and directly observe the mechanisms that generated it. When doing this, it is important to save the machine learning data and to note the conditions under which it was gathered. We include some guidelines as an appendix to this chapter.

There appear to be two outstanding limitations, however, to studying machine learning algorithms. The first problem is that they cannot provide abstractions or reflections about their behavior in a general way. While introspection is not a reliable data gathering technique, subjects' insights can be nevertheless helpful. It would be useful to have descriptions and summaries of the model's mental state, particularly when this is complex or time-based and changing. The second problem is that machine learning tends to be simple, done with a single type of data, a single knowledge base, a single learning algorithm, and a single and permanent learning goal (i.e., the machine is always and directly motivated). Although there are exceptions to each of these, in general these shortcomings often limit the application back to human learning and represent areas for further work for machine learning.

4.3 TYPES OF EXPERIMENTAL DESIGNS

We can outline several experimental designs for studying order effects. Note, however, that experimental designs are usually concerned with eliminating order effects by averaging over different sequences. More information on experimental design can be found in standard textbooks (e.g., Calfee, 1985; Campbell & Stanley, 1963).

4.3.1 Same tasks presented in different orders

The simplest design to study order effects is just to present the same task in different orders to different groups (between-groups design). Where this is possible, using simple direct measures of performance can detect whether different orders have an effect. With richer performance measures, such as verbal protocols, one might start to address how these different orders give rise to different behavior. Presenting the different orders to the same group (within-group design) is generally not applicable in learning studies, simply because subjects have learned after the first order!

It is also worth noting a direct and powerful trade-off: The more data you collect for a single subject, either by number of trials or by type of trials or by density of the data, the less

subjects that can be run for a given project size. With increased density or complexity of data more interesting questions can be answered. But the work then relies more on previous work (theories) to define how the initial results are to be interpreted, such as protocol theory, and the data gathering has more constraints on it. Furthermore, the analyses become more complex and difficult to perform.

An example from survey research helps illustrate the simplest form of a between-subject design for studying order effects. The example is concerned with a simple type of learning. Imagine a questionnaire with two items for assessing life happiness. Respondents have to indicate their happiness with life in general either before or after they have reported how happy they are with a specific domain of their life, namely dating. The dependent variable of interest is how dating accounts for general life happiness, which is indicated by the correlation between dating happiness and general happiness. The correlations will differ in the two conditions, with the correlation being higher when dating happiness is assessed after general happiness. The usual explanation for this order effect is that subjects interpret general life happiness as life happiness beside dating happiness when they where asked for dating happiness before (for a richer discussion of those issues see Strack, 1992).

Another example of order which is more closely related to learning is Asch's (1946) classical finding about impression formation in personal perception. Asch showed that a person will be viewed as more likeable when described as "intelligent-industrious-impulsive-critical-stubborn-envious" than when described by the same (!) list of traits presented in the opposite order. In Asch's view, this primacy effect occurs because some of the traits gain a positive meaning when preceded by a positive trait such as intelligent but gain a negative meaning when preceded by a negative trait such as envious. The trait adjectives thus seem to be differently colored in meaning depending on their context. This kind of experiment again uses a simple between-subjects design.

As another example of learning effects, consider industrial training programs that differ only in the order of presenting the material where subjects in different groups end up with different performance. Langley (1995) notes that sometimes the effects of two orders may converge with time (see the 'canceling out effect' in Figure 1, where further learning eventually cancels out the order effect seen in the first four stimuli). But keep in mind that the worse trained group will meanwhile be unnecessarily less productive until it catches up. This often matters!

4.3.2 Teaching sequences in educational psychology

In education psychology it is seldom possible (or meaningful) to carry out experiments where exactly the same information can be presented in different orders. Often in these experiments

the order and content cannot be completely independent variables, but order already results in some changes in the content. An example of this kind of design are the studies on effects of so called advanced organizers in learning texts (Griffin & Tulbert, 1995). Advanced organizers in the beginning of the text do not bring any additional information to the text but activate some prior knowledge and learning aims before reading. Exactly the same text or picture could be presented, for example, in the end of the text but this is not often the case in studies of the effects of advanced organizers.

On a more general level, sequence effects have been studied in comparison of different instructional approaches like discipline-based vs. problem-based methods in medical education (e.g., Schmidt, Machiels-Bongaerts, Cate, Venekamp, & Boshuizen, 1996). In principle, the same content can be taught but in very different orders. Discipline-based models start with the teaching of the theoretical basis of different medical disciplines. This knowledge will later be used in solving case problems, whereas the problem-based models start with authentic problems and the students have to study the basic science knowledge when solving these problems. From a methodological point of view, however, these two approaches almost never can be examined as a pure sequence effect because of the real-world limitations of presenting exactly the same material.

4.3.3 Observational methods for interacting with computers

In educational psychology many order sensitive experimental designs have been used where the learning sequence is not an independent variable but the dependent variable. Hypertext and instrumented readers (now called browsers) make it possible to follow students' personal reading sequences and problem solving strategies and then to compare the strategies with the learning results. Britt, Rouet, and Perfetti (1996) used hypertext documents and instrumented browsers as a method to present educational material. The format and the browser made it possible to record the students' reading sequences, which could then be analyzed with respect to learning and performance (see Scheiter & Gerjets, and Swaak & de Jong, this volume, for examples of this approach).

Another approach to work in this area when you have access to the user's computer is to use a keystroke logger such as RUI (Recording User Input, Kukreja, Stevenson, & Ritter, in press). Keystroke loggers record the user's keystrokes for later analysis and some, such as RUI, allow playback. This allows any piece of software to be studied. There are also commercial versions available that work with video data as well. Work under the topic of usability studies have further tools and methodological notes in this area.

4.3.4 Repeated surveys

There are areas where true experiments are not possible, where the situation cannot be independently manipulated. Repeated measurements through surveys and finding carefully matched cases can be used to measure longer term and more complex order effects in complex social behavior and education. For example, developmental psychologists have to note the order of development of skills; they can only survey the skill progression, not manipulate it. If they want to study the effects of the order of skill acquisition, they must observe many different, naturally occurring orders. The same elements will not always appear in each learning sequence (do all children hear the same words from their parents?), but we suspect that for many purposes they can be treated as equivalent on a more abstract level. Typically, the measurements brought forward to theoretical analysis are not the items themselves (such as performance on individual math problems), but higher level derived measures (such as a score on a standardized exam).

Surveys can also be done in the midst of another design. Subjects can be queried in the midst of performing a task with questions or requests or self-reports designed to measure their internal state in some way, for example, their motivational state (e.g., Feurzeig & Ritter, 1988). Scheiter and Gerjets (this volume) do this to study how students reorder problems on exams.

4.4 TYPES OF ANALYSES—STEPS TOWARDS THEORIES

How can those data on order effects be summarized? There are numerous examples of summary analyses in the other chapters, such as process models in psychology and machine learning algorithms in computer science. What we will address here are some of the preliminary analyses that need to be performed to understand the data prior to creating such models. These analyses can be used to summarize knowledge of order effects and to predict order effects.

4.4.1 Simple data descriptions

The first step in nearly any set of analyses is to create a set of simple descriptive statistics of the data, such as the response time for each task in each order and the number of errors per condition. It is often very useful to visualize the data in the form of a graph or a plot. This is part of a fairly large and well defined area of exploratory data analysis, for example, Tukey (1977) and Tufte (1990), which applies to sequential data as well. Sanderson and Fisher (1994) provide an overview of exploratory sequential data analysis (ESDA). Papers in their special issue on ESDA provide several examples (Frohlich, Drew, & Monk, 1994; Olson, Herbsleb, & Rueter, 1994; Ritter & Larkin, 1994; Vortac, Edwards, & Manning, 1994).

Simple descriptive statistical analyses are not always applicable when sequential data have been gathered. Keeping the sequential nature of the data precludes many averaging analyses, so creating graphic displays becomes more important. The user can try to create transition networks as behavioral summaries, such as Markov models, which show the frequency and types of transitions (Rauterberg, 1993; also see Sun & Giles, 1998 for more sequence learning models). More complex transition diagrams may include other features of the data such as the frequency of each category (Olson, Herbsleb, & Rueter, 1994).

Applying inferential statistics in this area can be a bit tricky. Often the assumptions of such analyses are violated by sequential data, such as independence. The best way to proceed is often to come up with a theory, and then simply work to improve it, not prove it (Grant, 1962).

4.4.2 Microgenetic analyses of verbal and other protocols

An extreme version of data analysis is to keep the sequential nature of the data completely intact, not using averages, but analyzing the data as a sequence of individual points. If simple reaction times are used as the data points, learning curves are generated because learning nearly always occurs (Ritter & Schooler, 2001).

Richer, non-numeric data are often kept as sequential data. Here, a sequential analysis of the data can extract more information and provide more direct information on how behavior is generated than reaction time means. This type of analysis includes protocol analysis (Ericsson & Simon, 1993; Newell & Simon, 1972) and microgenetic analysis (Agre & Shrager, 1990; Siegler, 1987; VanLehn, this volume). These analyses typically print or plot the data in series. The analyst then examines the data by hand looking for higher order patterns, such as strategies and strategy changes. This is a tedious form of analysis, but it often provides the most insight into behavior and its causes. The initial analyses are tied to data rather closely here, with the final analysis often designed to help create or test formal information-processing models.

4.4.3 Information processing process models

The analysis of order effects should lead to information processing theories that can make predictions about where, when, and how order effects will occur. These theories typically perform the task of interest, providing a description of the knowledge and mechanisms sufficient to perform a task. They typically will provide descriptions of intermediate internal states, the time course of processing, and near alternative actions. They do this in a very inspectable and objective way. This provides a much richer theory to test than a verbal theory.

These information processing models have been created in at least two traditions, machine learning, which emphasizes doing the task well, and cognitive modeling, which emphasizes doing the task like humans do the task. Summaries of theories are described in the other chapters on machine learning (Corneujols) and process models (Nerb et al.; Lane), and are used in many of the chapters (e.g., Gobet & Lane; Pavlik; Ohlsson). We will briefly preview them here and describe how they can influence data collection.

Process models make many predictions and many types of predictions. Nearly any type of data gathered can be compared with their performance. It is thus very easy to see where these types of theories are not matched by the data. Many people believe that this makes them bad theories. We believe that this viewpoint could not be more incorrect. If you are trying to create a theory that will predict behavior, you need to create a theory that makes strong predictions about behavior, which these theories do. Being able to see where the theory is not matched by the data allows you to improve or reject the theory. Theories that cannot be seen to be wrong, cannot be improved and, even worse, cannot be falsified. And, let us be fair, theories that do not make predictions are even more wrong for not making them. A more complete view of this theory development view is available from Grant (1962). Creating the model first thus points out what kinds of data to gather to validate and improve the model (Kieras, Wood, & Meyer, 1997).

4.5 CONCLUSIONS AND OPEN QUESTIONS

Because we cannot prove a theory, what is the role of data and what way forward do we have for organizing our understanding of order effects as theories? We believe that there are two, complementary ways. The first is simply to test your theories to show that they are worth taking seriously, and to find out where they are incomplete and could be improved (Grant, 1962). This approach does not end (how could it?), but is repeated until the theory is sufficient. The theory will remain incomplete and wrong in some way, for example, it will always be able to take account of further phenomena.

Project 1: Take your favorite task and design an experiment to study order effects. Can you augment the design of an existing experiment to look at order effects? Which experimental design would you use? What measurements would you take? What would taking different types of data mean? How would you analyze them? Would what you learn be worth the effort? If you are taking sequential measurements, how will you deal with the complexity? (See web sites for software and software reviews in this area of sequential data analyses.)

A second way forward is to start to create more broad theories. Here, the breadth of the theory counts as well as its depth. This approach, first argued for by Newell (1990), is for unified theories of cognition (UTCs). These theories are intended to include emotions, perception, and social interaction, so they might be better labeled unified theories of behavior. Practically, UTCs are currently studies on how to integrate theories, how to use a cognitive architecture, and on a host of practical problems. This approach is taken up in more detail in the introductory chapters that discuss models and the chapters on models in the second section of the book.

Project 2: How can we choose appropriate data to test our theories? Find a task in psychology, machine learning, education, or your own field. What are the typical measurements? What would be an unconventional measurement to take in your field that is routinely used in one of the other fields?

If there are theories or software to do so easily, run a pilot study creating and using this new technique for your area. An example of this would be to modify a machine learning algorithm or cognitive model (instead of a subject, typically) to "talk aloud" while it solves a task (this has only been done twice to our knowledge, Johnson, 1994, and Ohlsson, 1980).

Project 3: Can unified theories be correct? Does psychology need a uniform theory, or are first year undergraduates correct in saying that human behavior is too complex to understand let alone predict? What are some of the arguments for and against UTCs based on the data that is available? Prepare a 10 min. presentation.

REFERENCES

Agre, P. E., & Shrager, J. (1990). Routine evolution as the microgenetic basis of skill acquisition. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 694-701. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval. *Psychological Science*. *15*, 225-231.

Asch, S. E. (1946). Forming impression of personality, *Journal of Abnormal and Social Psychology, 41*, 303-314.

Baccino, T., & Kennedy, A. (1995). MICELAB: Spatial processing of mouse movement in Turbo-Pascal. *Behavior Research Methods, Instruments, & Computers, 27*(1), 76-78.

Belavkin, R. V., & Ritter, F. E. (2004). OPTIMIST: A new conflict resolution algorithm for ACT-R. In *Proceedings of the Sixth International Conference on Cognitive Modeling*, 40-45. Mahwah, NJ: Lawrence Erlbaum.

Byrne, M. D. (2001). ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies, 55*, 41-84.

Calfee, R. C. (1985). *Experimental methods in psychology*. New York, NY: Holt, Rinehart and Winston.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Clarke, D. D., & Crossland, J. (1985). *Action systems: An introduction to the analysis of complex behaviour*. London, UK: Methuen.

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*. Cambridge, MA: MIT Press.

Donchin, E., 1989. The Learning Strategies Project: Introductory remarks. Special Issue: The Learning Strategies Program: An examination of the strategies in skill acquisition, *Acta Psychologica 71*(1-3) 1-15.

Ericsson, K. A., & Simon, H. A. (1993). *Verbal protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Feurzeig, W., & Ritter, F. (1988). Understanding reflective problem solving. In J. Psotka, L. D. Massey, & S. A. Mutter (Eds.), *Intelligent tutoring systems: Lessons learned*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Frohlich, D., Drew, P., & Monk, A. (1994). Management of repair in human-computer interaction. *Human-Computer Interaction, 9*(3&4), 385-425.

Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers*. Cambridge, UK: Cambridge University Press.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*, 54-61.

Griffin, C. C. & Tulbert, B. L. (1995). The effect of graphic organizers on students' comprehension and recall of expository text: A review of the research and implications for practice. *Reading and Writing Quarterly, 11*. 73-89.

Hornof, A. J., & Halverson, T. (2003). Cognitive strategies and eye movements for searching hierarchical computer displays. In *ACM CHI 2003: Conference on Human Factors in Computing Systems*. 249-256. New York, NY: ACM.

Järvelä, S. (1996). *Cognitive apprenticeship model in a complex technology-based learning environment: Socioemotional processes in learning interaction*. Joensuu, Finland: Joensuu University Press.

Johnson, W. L. (1994). Agents that learn to explain themselves. In *The 12th National Conference on Artificial Intelligence (AAAI)*. 1257-1263. Menlo Park, CA: American Association for Artificial Intelligence.

Kibler, D., & Langley, P. (1988). Machine learning as an experimental science. In *Proceedings of the Third European Working Session on Learning*, (pp. 81-92). Glasgow: Pittman. Reprinted in J. W. Shavlik & T.G. Dietterich (Eds.) (1990), *Readings in machine learning*. San Francisco, CA: Morgan Kaufmann.

Kieras, D. E., Wood, S. D., & Meyer, D. E. (1997). Predictive engineering models based on the EPIC architecture for a multimodal high-performance human-computer interaction task. *Transactions on Computer-Human Interaction, 4*(3), 230-275.

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement. Vol. 1*. New York, NY: Academic Press.

Kukreja, U., Stevenson, W. E., & Ritter, F. E. (in press). RUI—Recording User Input from interfaces under Windows. *Behavior Research Methods, Instruments, and Computers*.

Langley, P. (1995). Order effects in incremental learning. In P. Reimann & H. Spada (Eds.), *Learning in humans and machines*. Kidlington, UK: Pergamon.

Lehtinen, E., Olkinuora, E., & Salonen, P. (Eds.). (1986). The research project on interactive formation of learning difficulties: Report III. A preliminary review of empirical results, Tom. B, No. 171. Turku, Finland: University of Turku.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Monty, R. A., & Senders, J. W. (Eds.). (1976). *Eye movements and psychological processes*. Hillsdale, NJ: Lawrence Erlbaum.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S. (1980). *Competence and strategy in reasoning with common spatial concepts: A study of problem solving in a semantically rich domain*. PhD thesis. Also published as #6 in the Working papers from the Cognitive seminar, Department of Psychology, U. of Stockholm, Stockholm.

Olson, G. M., Herbsleb, J. D., & Rueter, H. H. (1994). Characterizing the sequential structure of interactive behaviors through statistical and grammatical techniques. *Human-Computer Interaction*, *9*(3&4),427-472.

Oman, P. W., & Cook, C. R. (1990). Typographic style is more than cosmetic. *Communications of the ACM, 33*(5), 506-520.

Rauterberg, M. (1993). AMME: An automatic mental model evaluation to analyse user behavior traced in a finite, discrete state space. *Ergonomics, 36*(11), 1369-1380.

Rayner, K. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.

Ritter, F. E., Quigley, K. S., & Klein, L. C. (2005). Determining the number of model runs: Treating cognitive models as theories by not sampling their behavior. Unpublished mss.

Ritter, F. E., & Larkin, J. H. (1994). Using process models to summarize sequences of human actions. *Human-Computer Interaction, 9*(3&4), 345-383.

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In *International encyclopedia of the social and behavioral sciences*. 8602-8605. Amsterdam: Pergamon.

Ritter, F. E. (1988). Extending the Seibel-Soar Model. Presented at the Soar V Workshop held at CMU.

Roediger, R. (2004). What should they be called? *APS Observer, 17*(4), 5 & 46-48. Online: www.psychologicalscience.org/observer/getArticle.cfm?id=1549.

Salonen, P., Lehtinen, E. & Olkinuora, E. (1998). Expectations and beyond: The development of motivation and learning in a classroom context. In: J. Brophy (Ed.) *Advances in research on teaching. Vol. 7: Expectations in the classroom* (pp. 111-150). Greenwich, CT: JAI Press.

Salonen, P., & Louhenkilpi, T. (1989). Dynamic assessment of coping with failure, regulation of aspiration level, and comprehension skills. Study I: The effects of failure on the regulation of aspiration level. In M. Carretero, A. Lopez-Manjon, I. Pozo, J. Alonso-Tapia & A. Rosa (Eds.), *Third European Conference for Research on Learning and Instruction*. Madrid, Spain, September 4-7, 1989 (p. 95). Facultad Psicologia, Universidad Autonoma de Madrid.

Sanderson, P. M., & Fisher, C. A. (1994). Exploratory sequential data analysis: Foundations. *Human-Computer Interaction, 9*(3&4), 251-317.

Schmidt, H. G., Machiels-Bongaerts, M., Cate, T. J., Venekamp, R., & Boshuizen, H. P. (1996). The development of diagnostic competence: Comparison of a problem-based, an integrated, and a conventional medical curriculum. *Academic Medicine, 71*(6), 658-664.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology, 115*, 250-264.

Strack, F. (1992). Order effects in survey research: Activative and informative functions of preceding questions. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 23-34). New York, NY: Springer.

Sun, R., & Giles, C. L. (1998). *Sequence learning*. Berlin, D: Springer.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

VanLehn, K., Brown, J. S., & Greeno, J. (1984). Competitive argumentation in computational theories of cognition. In W. Kintsch, J. R. Miller & P. G. Polson (Eds.), *Methods and tactics in cognitive science* (pp. 235-262). Hillsdale, NJ: Lawrence Erlbaum.

Vortac, O. U., Edwards, M. B., Manning, C. A. (1994). Sequences of actions for individual and teams of air traffic controllers. *Human-Computer Interaction, 9*(3&4). 319-343.

APPENDIX: GUIDELINES FOR RUNNING MODELS AS SUBJECTS

Running models as subjects deserves some attention because it seems that everyone knows how to do it, but when it comes time to reexamine the model runs or reanalyze the results, problems appear. Here, we attempt to provide some guidance on how to run a model like a subject. The details will, of course, vary based on the size of the model and on the number of runs. For example, if the model simulates 90 hours of data, you might keep less than the full record of its behavior. If you are running the model 10,000 times, you might also not keep a full record of every run. Otherwise, it appears to us that model traces need to be treated as good as or better than empirical data.

There can be many intentions for running a model. One important reason is to understand your model, and another might be to illustrate how the model works so that you can explain how the mechanisms give rise to behavior (VanLehn, Brown, & Greeno, 1984). Another major intention of running a model is to generate predictions of behavior for comparison with data for validation and model development. And finally, an important reason is predicting human behavior. In this case you probably have run the model for comparison with human data to validate the model. At this point you may or may not want or need to compare it with human data.

In all of these cases, the predictions of the model should be clear. If your model is deterministic, then you only have to run your model once for a Soar model that learns. Some

Soar models are this way. If your model has stochastic (random) elements, you either need to compute what its expected value is using iterative equations (which we have only succeeded in doing once, Ritter, 1988), or you need to sample the model's behavior enough times that the model's predictions are clear.

Ideally, you would like a point prediction and a prediction of variance for each measure. Too often predictions are a sampled prediction, that is, the model's prediction is only an estimate of the model's final prediction because the model has not been run enough times.

Computing these predictions means running the model and saving its results. We can provide some suggestions for how to do this.

Suggestions

1. Save a copy of the model. "Freeze" it, so that at a later time the model, its cognitive architecture (if applicable), and any task apparatus that the model uses can be run or at least examined. This is similar to taking down study details like the experimental methods section of a paper for the model. Put this frozen copy in a separate directory from the model you are developing.

2. The model code should be documented to be at least as clear as good programs are. Dismal (www.gnu.org/software/dismal) provides an example that we can point to. Thus, model code should have an author, a date, preamble, required libraries and base systems, table of contents of the model, variables, and be presented as major sections. A README file should tell someone from another location how to load and run the model. This approach is based on a theory of how to write more readable code (Oman & Cook, 1990).

3. Record a trace of the model in enough detail that later analyses are possible. This is like recording individual differences and assigning a subject ID. It will thus be possible to run the model later, perhaps, if you need additional data. But if you are using a batch of runs, or the model takes a while to run, it will be very useful to have the longer trace available. If even larger traces are available, it is good insurance to record a few of these. These traces will also be helpful if at a later time you find something interesting in the model's behavior, or if you later want to report another aspect of the model's behavior. If you are running the model to represent different subject conditions, these traces should be separately recorded, labeled, and stored clearly.

4. Each run of the model, the trace and any summary measures, should be stored like subject data. That is, one run per file if possible, and not modified later.

5. The number of times to run a model is an interesting question. In nearly all cases, models are theories and as such their predictions should be and can be made very clear. Nearly all science theory we know of does not talk about sampling the theory, which is assumed to be fixed, but sampling data from the world. Thus, ideally the model should be run until its predictions are clear.

If your model is deterministic, running once is enough. If it has random components and is not deterministic, once is not enough. Increasing the number of model runs is nearly always much less expensive than increasing subjects. It is clear to us (Ritter, Klein, & Quigley, 2005) that there are several heuristics currently being used about how many times (e.g., "10" and "the number of subjects") that are not appropriate. Examining a limited number of subjects arises because of resource limitations; also they are data, and need to be treated differently. Model runs typically are not as limited.

A way to compute how many runs to perform is both necessary and possible. Looking at power and sample sizes have suggested to us that 100 runs will often provide fairly clear power for examining predictions for Cohen's medium (0.2) effect sizes. Power calculations will let you compute how many runs for a given effect size you would like to examine at a given confidence level of finding a difference. A 100 runs is currently more than most models are run. Power calculations suggest that 10,000 runs should nearly always be more than sufficient, but this can be problematic when the model takes a relatively long time to run.