Experimental Models for Validating Technology

Credits

Zelkowitz, M. V. and Wallace, D. R.
Experimental Models for Validating Technology
IEEE Computer 31, 5 (May. 1998), 23-31

Qualitative vs. Quantitative (1/4)

- What is that?
- What to prefer?
  - Are there strengths and weaknesses?
  - Is it influential the personality / thinking style?
- Might the one help the other?

Qualitative vs. Quantitative (2/4)

- **Qualitative research** involves analysis of data such as words (e.g., from interviews), pictures (e.g., video), or objects (e.g., an artifact).
- **Quantitative research** involves analysis of measurable attributes, eventually numerical data.
- **What to prefer?**
  - **The strengths and weaknesses** of qualitative and quantitative research are a perennial, hot debate, especially in the social sciences. The issues invoke classic 'paradigm war'.
  - **The personality / thinking style** of the researcher and/or the culture of the organization is under-recognized as a key factor in preferred choice of methods.
- **Qualitative helps in triangulating data**. Overly focusing on the debate of "qualitative *versus* quantitative" frames the methods in opposition. It is important to focus also on how the techniques can be integrated, such as in mixed methods research.

Qualitative vs. Quantitative (3/4) in Nine Points

1. "All research ultimately has a qualitative grounding"

   - *Donald Campbell* [Social Science]
2. The aim is a complete, detailed description.
3. Researcher may only know roughly in advance what s/he is looking for.
4. Recommended during earlier phases of research projects.

1. "There's no such thing as qualitative data. Everything is either 1 or 0”
   - *Fred Kerlinger* [Behavioral Science]
2. The aim is to classify features, count them, and construct statistical models.
3. Researcher knows clearly in advance what s/he is looking for.
4. Recommended during latter phases of research projects.

Qualitative vs. Quantitative (4/4) in Nine Points

5. Researcher is the data gathering instrument
6. Data is in the form of words, pictures or objects
7. Subjective - individuals' interpretation of events is important
8. Qualitative data is more 'rich', time consuming, and less able to be generalized.
9. Researcher tends to become subjectively immersed in the subject matter.

5. Researcher uses tools, such as questionnaires or equipment to collect numerical data.
6. Data is in the form of numbers and statistics.
7. Objective – seeks precise measurement & analysis of target concepts
8. Quantitative data is more efficient, able to test hypotheses.
9. Researcher tends to remain objectively separated from the subject matter.

Types

- An **historical method** collects data from projects that have already been completed [“**post-mortem**” study]. The data already exists; it is only necessary to analyze what has already been collected.
- An **observational method** will collect relevant data as a project develops. There is relatively *little control* over the development process [e.g., other than using the new technology that is being studied].
- A **controlled method** provides for multiple instances of an observation in order to provide for statistical validity of the results. This is the more classical method of experimental design in other scientific disciplines.

Historical Method
Literature Search

- The literature search represents the *least inva*sive (intrusion is null) and *most passive* form of data collection (data is collected whatever the project generates). It requires the investigator to analyze the results of papers and other documents that are publicly available.
- This inexpensive method places no demands on a given project and provides information across a broad range of domains.
- The major weakness with a literature search is selection bias or the tendency of researchers, authors, and journal editors to publish positive results. *Contradictory results often are not reported*, so a meta-analysis of previously published data may indicate an effect that is not really present if the full set of observable data was presented.

Historical Method
Study of Legacy Data

- We often want to understand a **previously completed project** in order to apply that information on a new project under development.
  - Investigated artifacts can include the source program, specification, design, and testing documentation, as well as data collected in its development (, i.e., the process).
- We assume there is a fair amount of quantitative data available for analysis.
  - When we do not have such quantitative data, we call the analysis a **lessons learned study** (described later).
  - We will also consider the special case of looking at source code and specification documents alone under the separate category of **static analysis** (also described later).

Historical Method
Lesson Learned

- Lessons-learned documents are often produced after a large industrial project is completed. A study of these documents often reveals *qualitative aspects* which can be used to improve future developments. If project personnel are still available, it is possible to *interview* them to understand the effects of methods used.
- Such data is severely limited.
- This form of project may indicate various trends, but cannot be used for statistical validity of the results.

Historical Method
Static Analysis

- We can often obtain needed information by looking at the completed product, which we call the static analysis method. This is a special case of studying legacy data except that we centralize our concerns on the *product* that was developed, whereas legacy data also included development process measurement. In these cases, we analyze the structure of the product to determine characteristics about it.
  - Software complexity and data flow research fit under this model. For example, since we do not fully understand what the effective measurements are, the assumption is made that products with a lower complexity or simple data flow will be more effective.

Observational Method
Project Monitoring

- Represents the lowest level of experimentation and measurement. It is the collection and storage of data that occurs during project development.
  (NdR. We do not have control on subjects assignment to treatments.)
  - [It is a *passive model* since] the available data will be collected whatever the project generates; intrusion is really limited. The assumption is made that the data will be used for some *immediate analysis*.
    - As already mentioned, if an experimental design is constructed **after** [such a] project is finished, [because we will be short of quantitative data], then we would call this an historical lessons learned study (see later).

Observational Method

Case-study (1/3)

- In a case study, a project is monitored and *data collected over time*.
  The project is often a large development and would be undertaken whether data was to be collected or not. With a relatively minimal addition to the costs to the project, valuable information can be obtained on the various *attributes* characterizing its development.
- This differs from the project monitoring method above in that *data collection is derived from a **specific goal** for the project*. A certain attribute is monitored (e.g., reliability, cost) and data is collected to measure that attribute.
- Similar data is often collected from a class of projects to build a *baseline* to represent the organization's standard process for software development.

Observational Method
Case-study (2/3)

- While project monitoring is considered passive, a case study is an active intrusive method because of the influence we may have on the development process itself (e.g., measurement intrusion into the process).
- The strength of this method is that the development is going to happen regardless of the needs to collect experimental data, so the only additional cost is the cost of measuring the development for specified attributes and collecting this data.
- There are many developments currently happening, so if the organization is attuned to the needs for experimentation and data collection, data from many projects can be amassed over a short period of time.

Observational Method Case-study (3/3)

- The weakness of this method is that each development is relatively unique, so it is not always possible to compare one development profile with another.
  Determining trends and statistical validity becomes difficult.

Observational Method
Assertion

- There are many examples where the developer of the technology is both the experimenter and the subject of the study. Sometimes this may be a preliminary test before a more formal validation of the effectiveness of the technology.
- The experiment is a weak example favoring the proposed technology over alternatives. As skeptical scientists, we would have to view these as potentially biased since the *goal* is not to understand the difference between two treatments, but to show that one particular treatment (the newly developed technology) is superior.

Observational Method
Field Study

- It is often desirable to compare several projects simultaneously.
- Since *a primary goal* is often not to perturb the activity under study, it is often impossible to collect all relevant data.
- An *outside group* will come and *monitor* the subject groups to collect the relevant information *without intruding the process*.
- This is related to the case study, but is less intrusive to the development process.
- This model best represents an organization that wishes to measure its development practices without changing the process to incorporate measurement.

Pilot, or Feasibility Study (1/2)

- A pilot, or feasibility study, is a study designed to test logistics and gather information prior to a larger study, in order to improve the latter's quality and efficiency. A pilot study can reveal deficiencies in the design of a proposed experiment (described later) or procedure and these can then be addressed before time and resources are expended on large scale studies. The pilot study may, however, provide vital information on the severity of proposed procedures or treatments.

Pilot, or Feasibility Study (2/2)

- The decision to conduct a pilot study prior to embarking on the main research project can be a difficult one for researchers. Sometimes it is tempting to omit this step, especially if the main study has been reasonably well planned. Constraints of time and a rush to get on with the main study are common reasons for passing over pilot work. However, this approach is risky, as no matter how thoughtfully a study has been planned, there are likely to be unforeseen difficulties.

Controlled Methods
Synthetic Environment Experiments
(or Controlled Experiments) (1/2)

This method is mostly used to:

- Confirming / disconfirming hypotheses and theories,
- Exploring relationships among data points that describe one variable or across multiple variables,
- Evaluating accuracy of models, or
- Validating measures [“measurement models”].

Controlled Methods. (2/2)
Synthetic Environment Experiments

Controlled experiments provide the highest level of formality, rigor, and control on measure. Specific guidelines are available for their conduction and documentation.

A precondition for conducting controlled experiments is a clear **hypothesis**; this guides the researchers in all the *steps* of the experiment design, including which *variable* to include in the design and how to *measure* them.

Controlled Methods.
Replicated Experiments (1/2)

- The main benefit of replication is that it helps mature software engineering knowledge by addressing both *internal validity* and *external validity* problems.
  - Regarding **internal validity**, replications aims to explore the range of conditions under which the experimental results still hold.
  - Regarding **external validity**, replications aim to support the hypothesis of the independence between the results and peculiarities of the study context.

Controlled Methods. Replicated Experiments (2/2)

- However, even with effectively specified laboratory packages, transfer of experimental know-how can still be difficult due to the existence of *tacit knowledge*.
- Additionally, a confirmation that the results are consistent only when researchers use exactly the same experimental design and materials does not support the hypothesis that those results will be repeatable in industry.
- In contrast, changes on subjects, settings, and materials would provide a higher level of validity to the empirical results.

Controlled Methods.
Dynamic Analysis

- The controlled methods we have so far discussed generally evaluate the development process. We can also look at controlled methods that execute the product itself. We call these dynamic analysis methods. Many instrument the given product by adding debugging or testing code in such a way that features of the product can be demonstrated and evaluated when the product is executed.
  - For example, a tool which counts the instances of certain features in the source program (e.g., number of if statements) would be a static analysis of the program, whereas a tool which executed the program to test its execution time would be a dynamic analysis method.

Controlled Methods. **Simulation** (1/2)

- Related to dynamic analysis is the concept of simulation. We can evaluate a technology by executing the product using a *model of the real environment*. In this case we hypothesize, or predict, how the real environment will react to the new technology. If we can model the behavior of the environment for certain variables, we often can ignore other harder-to-obtain variables and obtain results more readily using a simulated environment rather than real data.
- By ignoring extraneous variables, a simulation is often easier, faster, and less expensive to run than the full product in the real environment.

Controlled Methods. Simulation (2/2)

- The real weakness in a simulation is a lack of knowledge of how well the synthetic environment we have created models reality. Although we can easily obtain quantitative answers, we are never quite certain how relevant these values are to the problem we are trying to solve.

Which model to use (out of 12) (1/2)

1. *Project monitoring*. Observe the use of the new tool in a project and collect the usual accounting data from the project.
2. *Case study*. Use the new tool as part of new development. Collect data to determine if the developed product is easier to produce than similar projects in the past.
3. *Assertion*. Use the new tool to test a simple 100 line program to show that it finds all errors.
4. *Field study*. Distribute the new tool across several projects; collect data on the impact that the tool had.
5. *Literature search*. Find other published studies that analyze the behavior of similar tools.
6. *Dynamic analysis*. Execute a program with a new algorithm and compare its performance with the earlier version of the program.

Which model to use (out of 12) (2/2)

7. *Legacy data*. Find a previously-completed project that collected data on using the tool; analyze this data to see if tool was effective.
8. *Lessons learned*. Find a completed project that used this tool; Interview participants to see if tool had an impact on the project.
9. *Static analysis*. Use a control flow analysis tool to see if one design method results in fewer logic errors than another design method.
10. *Synthetic Environment Experiment*. Have 20 [*typical*] programmers spend two hours trying to debug a [*typical*] module, half using the new tool and half using other techniques.
11. *Replicated experiment*. Develop multiple instances of a module both using and not using the new tool; measure differences.
12. *Simulatio*n. Generate a set of data points randomly and then execute the tool and another tool to determine effectiveness in finding errors in a given module.