Chapter 18
Single-Case Mechanism Experiments

A **single-case mechanism experiment** is a test of a mechanism in a single object of study with a known architecture. The research goal is to describe and explain cause-effect behavior of the object of study. This can be used in implementation evaluation and problem investigation, where we do real-world research. It can also be used in validation research, where we test validation models. In this chapter we restrict ourselves to validation research, and in the checklist and examples the object of study is a validation model.

Single-case mechanism experiments are at the same time single-case causal experiments (Chap. 14, Abductive Inference Design). They investigate the effect of a difference of an independent variable X (e.g., angle of incidence) on a dependent variable Y (e.g., accuracy). But not all causal experiments are mechanism experiments. In a mechanism experiment, the researcher has access to the architecture of the object of study and explains the behavior of the object of study in terms of this architecture. This is not true of other kinds of causal experiments.

In this chapter, I use the phrase *mechanism experiment* to indicate single-case mechanism experiments. The description of the research context, research problem, and design of a mechanism experiment should be documented in an *experiment protocol*. Events during execution of the experiment and details of data analysis should be documented in a separate report, sometimes called an *experiment log*. In the rest of this chapter, we discuss how the checklist of Appendix B is applied to single-case mechanism experiments used in validation research.

18.1 Context

Table 18.1 gives the checklist for the research context, written from the point of view of the researcher preparing to do the research. Mechanism experiments can be done at any point in the engineering cycle. Researchers may evaluate

R.J. Wieringa, *Design Science Methodology for Information Systems and Software Engineering*, DOI 10.1007/978-3-662-43839-8_18

Table 18.1 The checklist for the research context. Initial questions to position your research, written from the point of view of the researcher preparing to do the research

1. Knowledge goal(s)
   - What do you want to know? Is this part of an implementation evaluation, a problem investigation, a survey of existing treatments, or a new technology validation?
2. Improvement goal(s)?
   - If there is a higher-level engineering cycle, what is the goal of that cycle?
   - If this is a curiosity-driven project, are there credible application scenarios for the project results?
3. Current knowledge
   - State of the knowledge in published scientific, technical and professional literature?
   - Available expert knowledge?
   - Why is your research needed? Do you want to add anything, e.g. confirm or falsify something?
   - Theoretical framework that you will use?

implementations, investigate problems, and study validation models by single-case mechanism experiments:

- ☐ In an implementation evaluation, a researcher may test a tool used by a manufacturer in its manufacturing design process, with the goal of analyzing the current architecture of the tool. An example was given in Chap. 5 on implementation evaluation and problem investigation (p. 47).
- ☐ In a problem investigation, a researcher may be interested in the architecture of a particular social network, and the curiosity-driven research goal may be to find out what groups or social actors are components of this network, what their capabilities are, and how they interact. To learn about a social network, the researcher may send messages to a network or may ask experimental subjects to perform actions in the network. This would be single-case research because it concerns one social network, and it would be mechanism research because the researcher can track and trace mechanisms of interaction in the network.
- ☐ In validation research, researchers may test algorithms in the laboratory in a simulated context or may do a serious game within an artificial project to investigate the effects of a new way of developing software.

In this chapter, we restrict our attention to mechanism experiments that are used in validation studies, and so in our examples, the *knowledge goal* of the research project is to validate new technology. The *improvement goal* in our examples is to develop some new technology. *Current knowledge* about a new technology may be based on earlier versions of the same technology and on earlier research in the new technology. We will use two examples in this chapter, one of which is our old friend the DOA algorithm:

- ☐ The direction of arrival estimation algorithm developed in the DOA project [5] was tested with a simulated array antenna and a wave arrival scenario. The first three checklist questions were answered as follows:

  (1) The knowledge goal was treatment validation.
  (2) The higher-level engineering goal was the development of an accurate estimation algorithm. The project was utility driven, with an industrial sponsor.

(3) The knowledge context consisted of a problem theory described in Chap. 5 on implementation evaluation and problem investigation (p. 44) and of design theories described in Chap. 7 on treatment validation (p. 63).

☐ Kumar et al. [1] describe a set of simulations of different organizational coordination mechanisms for scheduling medical tests of patients in a hospital. This is a problem of workflow and information system design, as the mechanisms differ in the order of activities, the allocation of tasks to actors, and the information flow between actors. Each coordination mechanism is an artifact whose properties need to be investigated. Four mechanisms in total were compared. The first three checklist questions were answered as follows:

(1) The knowledge goal was to learn which of these four mechanisms would produce the best combination of patient flow (time taken by the medical tests) and tardiness (time after due date that test result becomes available). This is a validation research goal.
(2) The research was utility driven, with the improvement goal to better utilize hospital resources.
(3) The knowledge context is general scheduling knowledge and domain-specific knowledge about patient scheduling in hospitals.

18.2 Research Problem

Table 18.2 gives the checklist for research problems. Because we restrict ourselves to validation research, the object of study consists of an artifact prototype interacting with a simulation of the context. The *conceptual framework* therefore will include the framework already developed when the artifact was designed. This framework may have to be extended with constructs and indicators needed to measure the performance of the validation model.

In validation research, the *knowledge questions* may be about different aspects of performance of the artifact in context:

- Effect questions: What effects are produced by the interaction between the artifact prototype and the simulated context? Why?
- Requirements satisfaction questions: Do the effects of the simulation satisfy requirements? Why (not)?
- Trade-off questions: What happens if the artifact architecture is changed? Why?
- Sensitivity questions: What happens if the context is changed? Why?

The *population* of validation research is not the set of similar validation models, but it is the set of all real-world instances of artifact × context. The validation model is investigated to learn something about real-world behavior and is not interesting in itself. The trade-off and sensitivity questions help to clarify the population predicate. For which classes of artifacts can we expect similar performance? In which class of contexts?

☐ (4) The conceptual framework of validation research in the DOA project [5] consists of a conceptual framework for signal reception, described earlier in Chap. 5 on implementation evaluation and problem investigation (p. 44), and of the conceptual framework for the DOA estimation algorithm, described earlier in Chap. 7 on treatment validation (p. 63).

(5) There are two groups of knowledge questions:

Table 18.2 The checklist for the research problem, written from the point of view of the researcher preparing to do the research

4. Conceptual framework
– Conceptual structures? Architectural structures, statistical structures?
– Chance models of random variables: Semantics of variables?
– Validity of the conceptual framework? Clarity of definitions, unambiguous application, avoidance of mono-operation and mono-method bias?
5. Knowledge questions
– Open (exploratory) or closed (hypothesis-testing) questions?
– Effect, satisfaction, trade-off or sensitivity questions?
– Descriptive or explanatory questions?
6. Population
– Population predicate? What is the architecture of the elements of the population? In which ways are all population elements similar to each other, and dissimilar to other elements?
– Chance models of random variables: Assumptions about distributions of variables?

- ☐ What is the execution time of one iteration of the DOA algorithms? Is it less or more than 7.7 ms? Why?
- ☐ What is the accuracy of the DOA estimations? Can they recognize angles of at least 1°?

(6) The intended population is the set of DOA estimation algorithms running in a satellite TV system in cars.

☐ (4) Kumar et al. [1] use general scheduling concepts taken from the operations research literature, such as *earliest due date, tardiness, slack*, and *patient flow time*. In this particular problem context, *tardiness* is a measure for how late a test is completed after its due date, *slack* is the difference between the due date of a test and the current date, and *patient flow time* is the time between release time of the earliest test of a patient (the earliest date at which the earliest test of the patient can be taken) and completion time of the last test for that patient.

(5) Kumar et al. do not state their knowledge questions, but they investigate tardiness and flow time for different coordination mechanisms. So apparently the knowledge questions are:

- What are the tardiness and flow time of patient test scheduling for each coordination mechanism? Why?

These are effect questions, and trade-offs are analyzed for the compared mechanisms.

(6) The population is not specified explicitly, but from the motivating introduction we can conclude that it is the set of all hospitals, defined by the following architecture: They consist of the so-called medical units providing medical care to patients, such as neurosurgery and cardiology, and of ancillaries such as radiology and the blood laboratory, which perform tests on patients as ordered by the medical units. Units and ancillaries have their own objectives, such as providing comfort to patients and using resources optimally. The compared coordination mechanisms allocate decisions about when to do a test variously to units, ancillaries, or a central coordinator and assume different kinds of information flows between units and ancillaries.

18.3 Research Design and Validation

The design of mechanism experiments requires decisions about the acquisition of validation models, sampling, treatment, and measurement. It also requires alignment of these decisions with the planned inferences from the data. The checklists for inference design are given in Sect. 18.4, and we illustrate their application to research design here.

18.3.1 Constructing the Validation Model

Table 18.3 gives the checklist for the OoS, which in validation research is a validation model, consisting of an artifact prototype and a model of the context. The artifact prototype is constructed by the researcher, and the model of the context may be constructed by the researcher too, or it may be acquired by the researcher in some other way. For example, the artifact prototype may run in a real-world context used as model of other real-world contexts. Before we discuss validity of a validation model, we look at the examples:

☐ (7.1) Two prototypes were made of the estimation algorithms in the DOA project [5], one in Matlab and one programmed in C on an experimental processor. In both cases, the simulated context consisted of sources that transmit waves, a uniform linear antenna array that receives waves, a beamsteering component that calculates time delays across antennas, and a beamforming component that composes the signal to be processed by the rest of the system. Simulations with Matlab were done with 5 wave sources located at angles of $-30°$, $-8°$, $0°$, $3°$, and $60°$ with respect to the vector orthogonal to the antenna array. Between simulations, the number of antennas, the signal-to-noise ratio, and the number of snapshots taken from the antennas was varied.

Table 18.3 The checklist for the object of study, written from the point of view of the researcher preparing to do the research. The OoS is a validation model

7.1 Acquisition of Objects of Study (validation models)
– How do you construct a validation model? What architecture should it have?
– Validity of OoS
- *Inference support.* Which inferences would be valid with respect to this design? See checklists for validity of descriptive statistics, abductive and analogic inferences.
- *Repeatability.* Could other researchers use your report to construct or select a similar OoS?
- *Ethics.* Are people informed that they will be studied, and do they consent to this? Are they free to stop at any time without giving reasons, and do they know this?

☐ (7.1) The artifacts validated by Kumar et al. [1] were coordination mechanisms between medical units and ancillaries to schedule patient tests:

- In a decentralized mechanism, each ancillary schedules patient tests independently.
- In a balanced coordination mechanism, the requesting medical unit imposes a patient flow due date.
- In a centralized mechanism, ancillaries additionally coordinate their schedules among each other.
- In a totally centralized mechanism, all coordination is done by a single scheduler [1, pp. 225–228].

These four artifacts were tested in a simulation of a hospital with three units, four ancillaries, and 30 patients [1, pp. 225].

To be valid, the validation model must support descriptive, abductive, and analogic inferences. Remember that we are using the term "validation" in two ways. In the engineering cycle, we assess the *validity of a treatment design* with respect to the problem it is designed for, and in the empirical cycle, we assess the *validity of inferences*. Here, we are interested in the second kind of validity. To increase the validity of inferences based on a validation model, the validation model must satisfy some requirements.

For *descriptive inference*, it is important that the chance model for variables is defined (Table 18.7). The meaning of indicators is defined in terms of observable properties of the validation model, i.e., of the artifact prototype and the model of the context. If symbolic data will be produced, then interpretation procedures have to be agreed on too.

Next, consider *abductive inferences* (Table 18.8). Validation models can support causal inference if the conditions for single-case causal experiments or comparative-cases causal experiment, listed in Chap. 14 (Abductive Inference Design), are satisfied. If the validation model contains people and you want to do causal inference, you have to assess possible threats to internal validity related to psychological or social mechanisms of people in the validation model or across validation models:

- *OoS dynamics*. Could there be interaction among validation models? Could there be interaction among people in a validation model? Could there be historical events, maturation, and dropout of people?

Whether or not we can do causal inference, we should also try to explain phenomena architecturally. To assess support *architectural inference,* the following questions are relevant:

- *Analysis*. Is there enough information about the architecture of the artifact and context available to do an interesting analysis later? Is the information exact enough to do a mathematical analysis? You may want to specify software, methods, techniques, etc., formally enough and list exactly the assumptions about entities and events in the context, to be able to do a precise analysis. This will also facilitate explanatory analysis of observed phenomena later on.

- *Variation*. What is the minimal validation model that you can construct to answer your knowledge questions? Can you omit components that have been specified in the population predicate? Can the components in a model that you constructed have restricted capabilities and still provide sufficient information to answer your knowledge questions? Which generalizations will be licensed by the similarity between the validation model and an artifact implemented in the real world? Varying the architecture of the artifact prototype, we actually do a trade-off analysis. Varying the architecture of the context simulation, we do a sensitivity analysis.
- *Abstraction*. The artifact prototype and context simulation will contain components not specified in the artifact design but required to run the simulation. Do these influence the behavior of the validation model? Will there be unwanted influences from parts of the model of the context that cannot be controlled? If we want to study the effect of mechanisms in their pure, undisturbed form, we should eliminate the influence of components and mechanisms not specified in the architecture. On the other hand, if we want to test the robustness of the architectural mechanisms under various disturbing influences, we should keep them.

A validation model may contain human actors, such as students who simulate a real-world software engineering project. In that case, you may want to be able to give *rational explanations* of observed behavior. If you want to do this, you must prepare for it by ensuring that you can get information about goals and motivations of actors in the simulation. The threats to validity of rational explanations are these:

- *Goals*. Actors may not have the goals that the explanation says it has. Can you get accurate information about the true goals of actors?
- *Motivation*. A goal may not motivate an actor as much as the explanation says it did. Can you get information about the motivation of actors in the simulation?

Generalization from a single-case experiment is done by architectural analogy. The following questions are important to assess support for *analogic inference* (Table 18.9):

- *Population predicate*. Will the validation model satisfy the population predicate? In which way will it be similar to implemented artifacts operating in a real-world context? In which way will it be dissimilar?
- *Ambiguity*. What class of implemented artifacts in real-world contexts could the validation model represent? What could be the target of analogic generalization?

In addition to supporting inferences, construction of the OoS must be repeatable and ethical. Other researchers must be able to *repeat* the construction, and if people are participating, the demands of *ethics* require that participants sign an informed consent form and must be informed that they are free to stop at any time:

☐ (7.1 continued) The architecture of the DOA validation model has been specified mathematically, and the expected effects produced by it can be proven mathematically. The model contains all information to provide architectural explanations of phenomena.

But the model abstracts from many components present in the real world. A car moving on a highway will pass many obstacles that may distort the waves, and these influences have been idealized away in the simulation. And the model idealizes away variation that is present in the real world: slightly unequal distances between antennas in an array, waves that are not quite plane, etc. Support for analogic generalization to the real world is not unlimited, and field tests in real-world contexts must be done to find out if these idealizations matter.

☐ (7.1 continued) The hospital architecture assumed by the study of Kumar et al. does not allow mathematical analysis, but it does allow analysis of events generated during a simulation. It contains sufficient information to give architectural explanations of phenomena.

But architectural components and capabilities in the model may have capabilities not present in their real-world counterparts. For example, the model assumes uncertainty about inputs, e.g., about due dates, and they do not claim effectiveness when inputs are uncertain, and it assumes that the hospital departments will not resist change [1, p. 235]. This limits generalizability to real-world hospitals.

In addition, real hospitals may contain components that are abstracted away in the model, such as independent practitioners working on the premises of the hospital and independent labs who do medical tests for the hospital. These additional components of real-world cases may disturb phenomena produced in a simulation.

18.3.2 Sampling

It may come as a surprise, but in single-case mechanism experiments, we sample objects of study too. In validation research, we construct a sample of validation models in sequence. As explained in Chap. 7 (Treatment Validation), this is a process of scaling up from the laboratory to the field, so the sequence of validation models studied starts in the lab and ends in the field. As in all processes of analytical induction, we construct confirming as well as disconfirming cases. Confirming validation models aim to replicate phenomena produced by earlier models; disconfirming models are extreme models used to explore the boundary of the conditions under which the phenomena can and cannot be produced.

The checklist for sampling is given in Table 18.4. The relevant validity consideration is the one for analogic inference:

- *Representative sampling*, case-based research. In what way will the constructed sample of models be representative of the population?

At the start of a process of scaling up to practice, our models are not representative of implemented artifacts in real-world contexts. They are tested under idealized conditions in the laboratory to assess feasibility of a design idea. Later, when we scale up to conditions of practice, the models become more realistic. During the process of scaling up, generalizability to the real world will become increasingly important, possibly supported by a theory of similitude.

Table 18.4 The part of the checklist for sampling objects of study for single-case mechanism experiments, written from the point of view of the researcher preparing to do the research. Objects of study are validation models

7.2 Construction of a sample
– What is the analytical induction strategy? Confirming cases, disconfirming cases, extreme cases?
– Validity of sampling procedure
- *Inference support.* Which inferences would be valid with respect to this design? See the applicable parts of the checklists for validity of statistical, abductive and analogic inferences.
- *Repeatability.* Can the sampling procedure be replicated by other researchers?
- *Ethics.* No new issues.

Table 18.5 Checklist for treatment design of a mechanism experiment in a validation study, written from the point of view of the researcher preparing to do the research. The OoS is a validation model. A treatment of a validation model is a scenario in which the context provides stimuli to the artifact

8. Treatment design
– Which treatment(s) will be applied?
– Which treatment instruments will be used? Instruction sheets, videos, lessons, software, computers, actuators, rooms, etc.
– How are treatments allocated to validation models?
* Are treatments scaled up in successive validation models?
– What is the treatment schedule?
– Validity of treatment design:
* *Inference support.* Which inferences would be valid with respect to this design? See the applicable parts of the checklists for validity of statistical, abductive and analogic inferences.
* *Repeatability:* Is the specification of the treatment and the allocation to validation models clear enough so that others could repeat it?
* *Ethics.* Is no harm done, and is everyone treated fairly? Will they be informed about the treatment before or after the study?

18.3.3 Treatment Design

Table 18.5 gives the checklist for designing treatments of a validation model. A validation model consists of an artifact prototype and a model of the context, and treatments are *scenarios* that the validation model is exposed to. There are some surprising confusions here that I will illustrate using a simple example of drug research. Consider a test of an experimental medicine, in which subjects are instructed to take the medicine according to some medical protocol, e.g., every morning before breakfast for the next 6 weeks. The medicine is the artifact; the

patient and his or her environment are the context. In this situation, there are three different interpretations of the term "treatment":

- The patient is treated by an experimental medicine. In other words, the context is treated by the artifact.
- An experimental medicine is tested by treating it to a realistic context. In other words, the artifact is treated by the context.
- The patient is instructed to take an experimental medicine according to a medical protocol. In other words, the artifact and context are treated to a scenario.

In validation research, we use the word "treatment" in the third sense. For example, an experimental software prototype may be treated to an input scenario from a simulated context, or in a serious game the participants may be treated to a scenario from a simulated context.

To deliver the treatment scenario, *treatment instruments* may be needed, such as software to generate scenarios, sensors to collect data from a simulated context, instruction sheets, videos or lessons for human participants, equipment and rooms to put them in, etc.

Treatments must be *allocated* to objects of study, which in validation research means that the researcher must decide which application scenarios to test on which models. When scaling up from lab to practice, the first models are exposed to toy scenarios, and the final models are exposed to realistic scenarios.

All of this must be *scheduled*. This is a practical and important matter because the schedule is limited by research budgets and research project deadlines.

For *causal inference*, the following questions are relevant to assess the degree of support of a treatment for causal explanations (Table 18.8):

- *Treatment control*. What other factors than the treatment could influence the validation models? If a validation model contains people, then possible influences are the treatment allocation mechanism, the experimental setup, the experimenters and their expectations, the novelty of the treatment, compensation by the researcher, and rivalry or demoralization among subjects. For software or hardware in the validation model, we would have to consider virtual or physical factors that could influence their behavior.
- *Treatment instrument validity*. If you use instruments to apply the scenario, do they have the effect on the validation model that you claim they have?

To conclude something about the target of the validation model, we do *analogic inference*. To assess support for analogic inference, the following questions are relevant (Table 18.9):

- *Treatment similarity*. Is the specified treatment scenario in the experiment similar to treatments in the population? Or are you doing an extreme case study and should it be dissimilar?
- *Compliance*. Is the treatment scenario implemented as specified?
- *Treatment control*. What other factors than the treatment could influence the validation models? This is the same question as mentioned above for causal

inference. The relevance for analogic generalization is that if there are factors that we could not control, we should ask if the implemented treatment should be interpreted as another treatment, namely, as the intended treatment plus uncontrolled factors.

Increased control over extraneous factors improves support for causal inference (internal validity) but decreases support for analogic inference to field conditions (external validity) because it makes the simulation less realistic.

In addition to support for inferences, treatment validity includes repeatability and ethics. The experiment protocol must specify the treatment scenarios explicitly, so that other researchers could *repeat* the test using their own validation model. These tests are replications: The results obtained earlier must be reproducible.

If people are involved, *ethical* considerations apply. People must be treated fairly, and no harm must be done. If deception is used, for example, by withholding some information from the subjects, this must not be unfair or harmful either. In a debriefing after the experiment, subjects must be informed of the true research goal, questions, design, and results of the experiment:

- ☐ (8) In the DOA test, the scenarios are all combinations of values for signal-to-noise ratios, numbers of snapshots, and number of antennas. No treatment instruments were needed other than the Matlab tool. The treatment scenarios were not intended to be fully similar to real-world scenarios, but they were realistic enough to be able to assess which algorithm was most promising in the intended context and could therefore be selected for further investigation. The researcher had full control of all factors that could influence the validation model. This improved support for causal inference but decreased support for analogic inference to real-world conditions.
- ☐ (8) Kumar et al. [1, p. 225] tested many scenarios in which six parameters were varied: number of tests per patient, test start times, processing time, due dates, load of ancillaries, and patient flow and test tardiness objectives. These are all representative of real-world scenarios. No treatment instruments were needed. The level of control was high, which in this example too improved support for causal inference but decreased support for analogic inference to the real world.

18.3.4 Measurement Design

Table 18.6 gives the checklist for measurement specification. Measurement requires the definition of *measured variables and scales*, and these are usually defined already in the conceptual framework of the research, which for validation research has already been designed as part of artifact design.

The *data sources* are components of the validation model from which you will acquire data, e.g., software or hardware components or people participating in the simulation. *Measurement instruments* include clocks, sensors, probes in software, log analyzers, as well as interviews and questionnaires for people participating in the experiment. There should be an infrastructure for *storing and managing measurement data*. Traceability of data to its source (provenance) and availability of the data to other researchers should be decided on.

Table 18.6 Checklist for measurement design of a mechanism experiment, written from the point of view of the researcher preparing to do the research

9. Measurement design
   - Variables and constructs to be measured? Scales, chance models.
   - Data sources? People (e.g. software engineers, maintainers, users, project managers, politically responsible persons, etc.), primary data (e.g. source code, log files, bug tracking data, version management data, email logs), primary documents (e.g. project reports, meeting minutes, organization charts, mission statements), etc.
   - Measurement instruments? Interview protocols, questionnaires, video recorders, sound recorders, clocks, sensors, database queries, log analyzers, etc.
   - What is the measurement schedule? Pretests, posttests? Cross-sectional or longitudinal?
   - How will measured data be stored and managed? Provenance, availability to other researchers?
   - Validity of measurement specification:
     * *Inference support.* Which inferences would be valid with respect to this design? See the applicable parts of the checklists for validity of abductive and analogic inferences.
     * *Repeatability.* Is the measurement specification clear enough so that others could repeat it?
     * *Ethics.* Which company data must be kept confidential? How is privacy of persons respected?

Validation models may support *causal inference* if they are used in single-case and comparative-cases causal experiments. But even if they are not used this way, the validity threats of causal inference are still relevant because you want to avoid disturbance of the validation model. The important question with regard to measurement is then the following (Table 18.8):

- *Measurement influence*. Will measurement influence the validation model?

If it does, then this should be subtracted from the data in order to identify treatment effects.

To assess support for generalization by *analogic inference*, the following questions must be answered (Table 18.9):

- *Construct validity*. Are the definitions of constructs to be measured valid? Clarity of definitions, unambiguous application, avoidance of mono-operation and mono-method bias?
- *Measurement instrument validity*. Do the measurement instruments measure what you claim that they measure?
- *Construct levels*. Will the measured range of values be representative of the population range of values?

Finally, measurements should be *repeatable* by other researchers and should be *ethical* for any human subjects. Confidentiality and privacy should be respected:

- The variables measured in the DOA project [5] are *degrees* of incidence and *decibels* of the signal. Their definitions are taken from the literature and need not be included in a research report.

  The artifact prototype itself and the simulation of the context are at once the source of data and the measurement instruments.

  The researcher took care to spread the angles of incidence, signal-to-noise ratios, and numbers of antennas to have a reasonable coverage of the ranges of these values in real-world situations. This reduces the construct level validity threat mentioned above. Practical aspects of research design, such as data storage and management, are not reported.
- Kumar et al. [1] measure variables like *patient blocked time*, *test processing time*, and *ancillary blocked time*. These are defined in their conceptual research framework, and the authors provide arguments toward the validity of these variables as indicators of the efficiency of the coordination mechanisms studied.

  The simulation software at once generates the simulations, is the source of data, and is the measurement instrument. Parameters of the simulation are *patient load*, *mean inter-arrival time*, *due date*, etc., and these were set to values for a small hospital [1, p. 225]. This introduces a construct level validity threat. Practical aspects of research design, such as data storage and management, are not reported.

18.4 Inference Design and Validation

Single-case mechanism experiments are case based, and inferences from them are done in three steps: description, architectural explanation, and generalization by analogy. The construction of validation models is done with the goal of supporting these kinds of inference, and validity considerations for them have already been given above. Examples of the inferences themselves are given later, in the section on data analysis. Here we have a brief look at the relevant parts of the checklist.

Table 18.7 gives the checklist for descriptive inference. Descriptive inference in single-case mechanism experiments is often the presentation of data in digestible form such as graphs or tables with aggregate information. As usual, data may be transformed, and symbolic data such as images or text must be interpreted. The validity requirements for descriptive inference all ask in one way or another whether the researcher added information to the data that is not warranted by the observed phenomena or by prior knowledge.

Table 18.8 gives the checklist for abductive inference. If the behavior of the validation model is time independent and if effects are transient, then you can do single-case causal experiments with them. And if they can be replicated, you can do comparative-cases causal experiments with them. Since the architecture of the validation model is known, you should try to explain causal relations established this way architecturally. If the validation model contains people, then you may be able to explain their behavior rationally.

Table 18.9 gives the checklist for analogic inference. Generalization from mechanism experiments is done by architectural analogy: In objects with a similar architecture, similar mechanisms will produce similar phenomena. The purpose of experimenting with validation models is to assess the required similarity between model and target. This goes both ways: How similar must a validation model

Table 18.7 Checklist for descriptive inference, written from the point of view of the researcher preparing to do the research. The OoS is a validation model

10.1 Descriptive inference design

- How are words and images to be interpreted? (Content analysis, conversation analysis, discourse analysis, analysis software, etc.)
- What descriptive summaries of data are planned? Illustrative data, graphical summaries, descriptive statistics, etc.
- Validity of description design
  - *Support for data preparation.*
    - Will the prepared data represent the same phenomena as the unprepared data?
    - If data may be removed, would this be defensible beyond reasonable doubt?
    - Would your scientific opponents produce the same descriptions from the data?
  - *Support for data interpretation.*
    - Will the interpretations that you produce be facts in your conceptual research framework? Would your scientific peers produce the same interpretations?
    - Will the interpretations that you produce be facts in the conceptual framework of the subjects? Would subjects accept them as facts?
  - *Support for descriptive statistics.*
    - Is the chance model of the variables of interest defined in terms of the population elements?
  - *Repeatability:* Will the analysis repeatable by others?
  - *Ethics:* No new issues.

be to real-world implementations to learn something from the model about those implementations? Conversely, how similar must an implementation be to show behavior similar to the validation model? You test models that represent different artifact versions and with variations of the context. Testing prototypes of different artifacts in the same context is trade-off analysis, and testing different contexts with the same artifact is sensitivity analysis. If the mechanism experiment is part of a process of scaling up from the laboratory to the field, then at every step, the research goal is to acquire sufficient certainty about the repeatability of behavior at the current scale, so as to justify the step to the next level of scaling up.

Table 18.8 Checklist for inference design of a mechanism experiment in validation research, written from the point of view of the researcher preparing to do the research. The OoS is a validation model

10.3 Abductive inference design

- What possible explanations can you foresee? What data do you need to give those explanations? What theoretical framework?
- Internal validity
  - *Causal inference*
    - *OoS dynamics.* Could there be interaction among OoS's? Could there be historical events, maturation, and drop-out of OoS's?
    - *Treatment control.* What other factors than the treatment could influence the validation models? The treatment allocation mechanism, the experimental setup, the experimenters and their expectations, the novelty of the treatment, compensation by the researcher, rivalry or demoralization about the allocation?
    - *Treatment instrument validity.* If you use instruments to apply the scenario, do they have the effect on the validation model that you claim they have?
    - *Measurement influence.* Will measurement influence the validation models?
  - *Architectural inference*
    - *Analysis:* The analysis of the architecture may not support its conclusions with mathematical certainty. Components fully specified? Interactions fully specified?
    - *Variation:* Do the real-world case components match the architectural components? Do they have the same capabilities? Are all architectural components present in the real-world case?
    - *Abstraction:* Does the architectural model used for explanation omit relevant elements of real-world cases? Are the mechanisms in the architectural model interfered with by other mechanisms, absent from the model but present in the real world case?
  - *Rational inference*
    - *Goals.* An actor may not have the goals assumed by an explanation. Can you get information about the true goals of actors?
    - *Motivation.* A goal may not motivate an actor as much as assumed by an explanation. Can you get information about the true motivations of actors?

Table 18.9 Checklist for inference design of a mechanism experiment in validation research, written from the point of view of the researcher preparing to do the research. The OoS is a validation model

10.4 Analogic inference design

- What is the intended scope of your generalization?
- External validity
  - *Object of Study similarity.*
    - *Population predicate.* Will the validation model satisfy the population predicate? In which way will it be similar to implemented artifacts operating in a real-world context? In which way will it be dissimilar?
    - *Ambiguity.* What class of implemented artifacts in real-world contexts could the validation model represent? What could be the target of analogic generalization?
  - *Representative sampling,* case-based research: In what way will the constructed sample of models be representative of the population?
  - *Treatment.*
    - *Treatment similarity.* Is the specified treatment scenario in the experiment similar to treatments in the population?
    - *Compliance.* Is the treatment scenario implemented as specified?
    - *Treatment control.* What other factors than the treatment could influence the validation models? Could the implemented treatment be interpreted as another treatment?
  - *Measurement.*
    - *Construct validity.* Are the definitions of constructs to be measured valid? Clarity of definitions, unambiguous application, avoidance of mono-operation and mono-method bias?
    - *Measurement instrument validity.* Do the measurement instruments measure what you claim that they measure?
    - *Construct levels.* Will the measured range of values be representative of the population range of values?

Table 18.10 The part of the checklist for research execution relevant for mechanism experiments, written from the point of view of the researcher preparing to write a report about the research

11. What has happened?
– What has happened when the OoS's were selected or constructed? Did they have the architecture that was planned during research design? Unexpected events for OoS's during the study?
– What has happened during sample construction? Could you build or acquire all objects of study that you planned to study?
– What has happened when the treatment(s) were applied? Mistakes, unexpected events?
– What has happened during measurement? Data sources actually used, response rates?

18.5 Research Execution

We now switch perspective from designing your research to executing it and reporting about it. Collecting a report starts as soon as you start executing the experiment. Table 18.10 lists the checklist items for reporting about research execution. Not everything that happens during execution of a research design needs to be reported. What information about events during research execution did you use to interpret your results? What information would be useful to provide if someone wants to repeat your research? The reader of a report must trust that the writer included all relevant information, so as a writer you will have to be honest:

☐ (11) The report about the DOA project gives no information about the Matlab models. However, it contains detailed information about the construction of the C program that implemented the MUSIC algorithm and the relevant properties of the experimental Montium2 processor on which it was executed [5, Chap. 5].

☐ (11) Kumar et al. [1] give no information about the construction of their simulation or about the events during simulation.

18.6 Data Analysis

We now perform the inferences planned, point no: 13 is missing. Please check and provide the same. for in our research design. Table 18.11 shows the checklist for data analysis. The part about statistical inference is absent because we are studying single cases, e.g., single simulations and single prototypes, and not samples of cases.

In the rest of this section, we give examples without further comments. I should repeat here that written reports may present information differently. For example, validity has been considered during research design and inference design and should

Table 18.11 The part of the checklist for data analysis that is relevant for a mechanism experiment, written from the point of view of the researcher preparing to write a report

12. Descriptions
    - Data preparations applied? Data transformations, missing values, removal of outliers? Data management, data availability.
    - Data interpretations? Coding procedures, interpretation methods?
    - Descriptive statistics. Demographics, sample mean and variance? Graphics, tables.
    - Validity of the descriptions: See checklist for the validity of descriptive inference.
14. Explanations
    - What explanations (causal, architectural, rational) exist for the observations?
    - Internal validity: See checklist for the validity of abductive inference.
15. Generalizations
    - Would the explanations be valid in similar cases or populations too?
    - External validity: See checklist for the validity of analogic inference
16. Answers
    - What are the answers to the knowledge questions? Summary of conclusions, support for and limitations of conclusions.

be reviewed again during data analysis. A written report may present the result of all validity considerations only once, for example, in a separate section; it may distribute the discussion over the different parts of research design, as we did here; or it may distribute the discussion over the different parts of the data analysis.

18.6.1 Descriptions

- ☐ (12) Vrielink [5] reports sensitivity of a 16-element antenna in different directions, the spectrum recognized by DOA algorithms in different directions, and DOA estimation errors in different directions. Execution times on the intended Montium processor were estimated, not observed, because the processor was not implemented yet. There are in addition qualitative observations such as that changing the number of antennas does not result in a significant difference in performance between the different estimation algorithms [5, p. 24]. This is a sensitivity property.
- ☐ (12) Kumar et al. [1] reported the percentage improvement (reduction) in tardiness of the different mechanisms over the base configuration of decentralized ancillaries. All improvement data are given in an appendix, and the paper contains numerous graphs visualizing the improvement trends for different settings of the parameters.

18.6.2 Explanations

- ☐ (14) The functional correctness of the output of the tested algorithms in the DOA project is explained by the algorithm structure. This is what the algorithms were designed for. The time performance properties are explained by the computational complexity of various steps [5, pp. 20, 55, 58]. He explained measured accuracy of different algorithms in terms of physical properties of the waves [5, p. 23]. These are architectural explanations.

  There are also causal explanations. For example, the number of antennas is positively correlated with the spatial resolution of the tested algorithms, and this is interpreted causally: Increasing the number of antenna causes an increase in spatial resolution of the compared algorithms [5, p. 24]. Presumably, this causal explanation in turn can be explained architecturally by the structure of the algorithms and architectural properties of the antenna-wave system.

  Other observations remain unexplained, such as that one algorithm performed better than the other in tests where the signal-to-noise ratio was low [5, p. 27].
- ☐ (14) As explained earlier, Kumar et al. studied four coordination mechanisms, a decentralized, balanced, centralized, and totally centralized one. See the discussion of item (7.1) above. The data showed that the totally centralized solution improved on the centralized ancillary solution, which improved on the balanced solution, which improved on the decentralized ancillary solution. The authors explain this by the fact that each of these mechanisms includes those that follow it in this list. Note that the quantitative amount of the observed improvements cannot be explained, but their ordering can.

18.6.3 Analogic Generalizations

- ☐ (15) The results of the DOA experiments are generalized to real-world implementations of the algorithms, running in a satellite TV system that is part of a car driving on a road. This generalization is supported by the similarity of algorithms in the laboratory and in the field. The laboratory simulation of the context may be less similar to the real-world context, because in the real world, various conditions of practice may disturb the results obtained in the laboratory simulation. Field tests are needed to give more support to the generalization.
- ☐ (15) The simulations by Kumar et al. can probably be replicated by other researchers, in which case they are generalizable, by architectural analogy, to other simulations. Generalization to real hospitals is less well supported. The simulation ignores uncertainty about inputs and resistance to change. There are many architectural capabilities, limitations, and mechanisms in a real hospital that may interfere with the simulated coordination mechanisms in a way that makes the results unreproducible in the real world. To learn more about this, real-world case studies should be done. If a hospital decides to implement one of these coordination mechanisms, then we may be able to study the resulting mechanisms in detail in this case. This would be an evaluation study [2].

Some generalizations may not be based on architectural similarity but on feature-based similarity. If we think that an observed phenomenon generalizes to similar cases without understanding the mechanism behind it, then we have postulated an **empirical regularity.** The phenomenon may indeed be regular, so that we can use it as a prediction. But if it remains unexplained, we should treat it as an empirical regularity that may be broken for reasons that we do not understand.

When practitioners use an empirical regularity for which we have no architectural explanation, they can manage the risk that the regularity is violated by moving in small steps. Designs that have been proven to work in practice are used in new situations with only small differences from proven cases. This will make the set of proven cases expand gradually. This is one of the motivations behind evolutionary design [3, 4]:

- For example, the observation in the DOA project that one algorithm performed better than the other in tests where the signal-to-noise ratio was low [5, p. 27] is unexplained. Suppose that this phenomenon cannot be explained in terms of the different structures of the algorithms. Then we can still treat it as an empirical regularity. There is a slight architectural flavor to this generalization, because the generalization is that in cases with similar architecture, the same phenomenon will occur. But as long as we do not understand how this architecture produces this phenomenon, we should treat it as an empirical regularity. When it is used, it is safe to use it in situations that only differ incrementally from situations where it has shown to be true.

18.6.4 Answers to Knowledge Questions

- (16) The DOA project has a number of effect questions:
  - What is the execution time of one iteration of the DOA algorithm? Is it less or more than 7.7 ms? Why?
  - What is the accuracy of the DOA estimations? Can they recognize angles of at least 1°?

  The data analysis provides answers to these questions for two algorithms that can be tentatively generalized from the laboratory to the field. One of the algorithms was shown to satisfy the requirements on execution speed and accuracy.
- (16) Kumar et al. did not state a knowledge question, but we assumed it to be the following:
  - What are the tardiness and flow time of patient test scheduling for each coordination mechanism? Why?

  The data analysis provided support for the generalization that in the absence of disturbing mechanisms and for a simulation of a small hospital, the four coordination mechanisms increasingly reduce tardiness, where total centralization gave the best results. Without further research, this cannot be generalized to real hospitals.

18.7 Implications for Context

Table 18.12 lists the checklist for relating the research results to the knowledge context and improvement context. The knowledge context in validation research is what is known so far about the artifact being validated. The improvement context is the engineering cycle of the artifact:

Table 18.12 The checklist for implications of the research results

17. Contribution to knowledge goal(s) Refer back to items 1 and 3.
18. Contribution to improvement goal(s)? Refer back to item 2.
    - If there is no improvement goal: is there a potential contribution to practice?

☐ (17) The DOA mechanism experiment [5] added knowledge about the performance of the estimation algorithms in the particular context of satellite TV reception.
(18) The research was sponsored by a hardware manufacturer, who used the results in the design and development of a new satellite TV system for cars.

☐ (17) Kumar et al. [1] were the first to apply their centralized and decentralized coordination mechanisms to patient test scheduling. So this was the first knowledge about these artifacts in this context.
(18) Their research was motivated by one particular hospital, and each of their mechanisms improved patient test scheduling performance compared to current performance in the hospital. The paper does not report if one of these mechanisms was actually adopted by the hospital.

References

1. A. Kumar, P.S. Ow, M.J. Prietula, Organizational simulation and information systems design: an operational level example. Manag. Sci. **39**(2), 218–240 (1993)
2. R. Pawson, N. Tilley, *Realistic Evaluation* (Sage, Thousand Oaks, 1997)
3. H. Petroski, *To Engineer is Human: The Role of Failure in Successful Design* (Vintage books, New York, 1982)
4. H. Petroski, *The Evolution of Useful Things* (Vintage Books, New York, 1992)
5. J.D. Vrielink, Phased array processing: direction of arrival estimation on reconfigurable hardware. Master's Thesis, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, January 2009. http://essay.utwente.nl/62065/