IRM-SA Experiment Plan

1 Scoping phase

- **Object of study** – what is studied?
The Invariant Refinement Method (IRM-SA) – both the process and the associated model.
- **Purpose** – what is the intention?
To evaluate IRM-SA in the design of DEECo-based systems
- **Quality focus** – which effect is studied?
The effectiveness of the IRM-SA design process
- **Perspective** – whose view?
The researcher’s point of view
- **Context** – where is the study conducted?
The experiment is run using M.Sc. students as subjects based on a defined lab exercise (object) of developing a non-trivial emergency coordination system for firefighters. The study is conducted as a “multi-test within object study”, since it involves many subjects (participants) and only one object (one example/system to be designed).

According to the template from [1], the goal of this empirical study (controlled experiment) is to:

```
Analyze the IRM-SA for the purpose of evaluation with respect to effectiveness from the point of view of the researcher in the context of M.Sc. students producing DEECo artifacts and linking them to requirement artifacts.
```

2 Planning phase

2.1 Context Selection
We will conduct an off-line (opp. on-line) experiment with students (opp. professionals) (see [2] for generalizing results from studies with students as subjects), addressing toy (opp. real) problems, in a general (opp. specific, e.g., bound to a specific company/environment) frame.

2.2 Hypothesis Formulation
- Null Hypothesis \( H_{01} \) = “Using IRM-SA to design a DEECo-based system does not result into a more effective design than using no systematic method.”
- Alternative Hypothesis 1 \( H_{11} \) = “Using IRM-SA to design DEECo-based system results into a more effective design than using no systematic method.”
2.3 Variables Selection

2.3.1 Independent variables

1. Design Method for DEECo-based systems
   a. Range: 2 values
      i. IRM-SA (Invariant Refinement Method for Self-Adaptivity)
      ii. NSM (No Systematic Method)
   b. Scale: Nominal

2.3.2 Dependent variables

1. Effectiveness of design process
   a. Range: natural numbers in [0..100] (worst/best)
   b. Scale: Ratio
   c. Directly measurable: No, we are measuring the quality of the final design artifact instead.
   d. Measurement method: Blind reviews and evaluation of final artifacts (designs in paper) produced the subjects (students) and performed by experts (us).
   e. Assumptions that the measurement method is based on:
      i. Reviewers are experts and not biased.
      ii. Subjects are similar in terms of background, curriculum, age, working experience, eagerness to participate, etc. (This way they won’t introduce any confounding factor to the experiment, i.e., their performance will be affected by the employed method more than by any other factor.)
      iii. The better the final artifact, the more effective the design process that leaded to the artifact.

2. Applicability of experiment settings, intuitiveness of IRM-SA concepts, etc.
   b. Scale: Ordinal
   c. Directly measurable: Yes
   d. Measurement method: Questionnaires
   e. Assumptions that the measurement method is based on: Subjects are not biased to answer in favor of IRM-SA.

2.4 Selection of Subjects

- Target population: Junior (Inexperienced) Software Engineers
- Chosen sampling technique: Convenience sampling, as the nearest and most convenient persons are chosen. Other techniques include:
  - Probability sampling
    - Simple random
    - Systematic
    - Stratified random
  - Non-probability sampling
    - Convenience
    - Quota
- Sample size: 20
  - The larger variability in the population, the larger size is needed.
2.5 Experiment Design

Multi-test within object study (one object, many subjects → many “runs”)

- Randomization
  - In the assignment of students to each treatment
  - In the selection from the available students ← only if there are many students available

- Blocking
  - By experience with distributed systems/general development ← only if there are many students available

- Balancing
  - In the number of students per group: The control and treatment groups have the same number of students (balanced design).
    - This simplifies and strengthens the statistical analysis of the data.

Standard design for “one factor with two treatments”

- Completely randomized design (opp. paired comparison/crossover design): comparing two treatment means

2.6 Instrumentation

- Experiment objects
  - Firefighters scenario specification & detailed requirements
  - Example of a final artifact – DEECo specification and linkage to individual requirements on a separate domain (e-cars)

- Guidelines
  - Guidelines and tutorial for use of IRM-SA
  - Checklist to be used for both IRM-SA and NSM users

- Measurement and data collection instruments
  - Manual forms

2.7 Validity Evaluation
Green: will address, and we already know how
Blue: will address, still not entirely clear how
Red: have to address, but it's not clear how
Underlined: will not address, but have to report and explain our decision
Bold: will address, and maybe it's good to point out how we did it
Stricken out: don't have to address, not applicable

1. Conclusion validity: threats that limit our ability to draw the correct conclusion about relations between the treatment and the outcome.
   i. Low statistical power
      1. Usage of a sample that is big enough (which is “big enough” in our case?)
   ii. Violated assumptions of statistical tests
      1. Our samples that are normally distributed – students with normally distributed characteristics, e.g., competence in programming (make a questionnaire to ensure that?)
      2. Use samples that are independent – in our case all samples are independent.
   iii. Fishing and the error rate
      1. Calculate the actual significance
   iv. Reliability of measures
      1. Use accurate question wording
      2. Use objective measures in the reviews (e.g. number of faults)
      3. Make the evaluation criteria, procedure and formula available (so that tests can be replicated with accuracy if needed)
   v. Reliability of treatment implementation
      1. Treat control and treatment team in the same way, e.g., spend equal amount of time explaining the scenario with both teams.
      2. Inevitably we will spend more time with the treatment group because we will have to explain also IRM-SA to them.
      3. Treat the two experiment blocks (CITY/D3S) in the same way, e.g., provide the same task, materials and guidelines.
      4. Run the experiment in standard working days (not close before/after holidays or special occasions/dates).
   vi. Random irrelevancies in experimental setting
      1. Take care that situations such as sudden interrupts (e.g. classroom not booked for the whole amount of time), too much noise, etc., won't arise during the experiment.
   vii. Random heterogeneity of subjects
1. **Students will be chosen from the same course and with similar curriculums, so they won’t be too heterogeneous. (However this has to be reported as a threat to external validity.)**

2. **Internal validity:** threats that can affect the independent variable with respect to causality, without the researcher’s knowledge. They threat the conclusion about a possible causal relationship between treatment and outcome.
   a. Single group threats
      i. **History**
         1. Schedule experiment occasions in “normal” days, i.e., not right before/after holidays, etc.
      ii. **Maturation**
         1. Do not make the experiment too long (students will be bored and demotivated if it's more than 1.5 hour)
         2. Take care that there are no learning (side-) effects like hints (or provide the same hints to all groups).
         3. **The design of the experiment (not crossover)** prevents learning effects.
      iii. **Testing**
         1. Do not repeat the test – we will only run the test once per subject.
      iv. **Instrumentation**
         1. Review and iterate over the example and the guidelines
         2. Test the example and guidelines in 1-2 researchers (acting as subjects) to obtain feedback and improve them before the actual experiment
   v. **Statistical regression**
      1. There is no previous experiment/case study
   vi. **Selection**
      1. Convenience selection ← weak point (but can we do anything about it? We don’t have a large pool of students.)
   vii. **Mortality**
      1. Monitor and characterize the dropouts from the experiment in order to check if they are representatives of the total sample (but we don’t expect to have many dropouts anyway).
   viii. **Ambiguity about direction of causal influence**
      1. Only one direction is possible, as the effectiveness cannot influence the choice of method.

b. Multiple group threats
   i. **Interactions with selection**
      1. Even though we will assign students to groups at random, we cannot eliminate the probability of one group maturing/learning faster than the other one, etc. These threats are referred to as "selection-maturation", "selection-history", etc.
c. Social threats
   i. Diffusion or imitation of treatments
      1. Make sure that the two groups do not talk to each other, so that the control group is not influenced by the treatment group, and vice versa.
      2. Better schedule the experiment for both teams on the same day to avoid students chatting with each other.
      3. There will be no experts in the area of DEECO-based design among the subjects, so no threat of specialized expertise.
   ii. Compensatory equalization of treatments
      1. Do not teach the control group another method as a substitute for not teaching them IRM-SA (this can influence the results a lot), but stress they have to use their intuition.
   iii. Compensatory rivalry
      1. In general, motivate the students to perform well but do not turn the exercise into a race.
      2. Do not inform the control group about the presence of IRM-SA, or
      3. Do not inform the control group what the treatment group is about to do (that the other ones will use IRM-SA)
      4. Do not inform the treatment group about the presence of the control group, or
      5. Do not inform the treatment group about what the control group is about to do (that the other ones won't use the IRM-SA)
   iv. Resentful demoralization
      1. Do not inform the control group about the presence of IRM-SA, or
      2. Do not inform the control group what the treatment group is about to do (that the other ones will use IRM-SA)

3. Construct validity: threats that limit our ability to generalize the result of our experiment to the concept or theory behind the experiment.
   a. Design threats
      i. Inadequate preoperational explication of constructs
         1. Clarify underlying theory of the experiment – it's not enough to say that we want to check if IRM-SA is 'better' – what does 'better' mean?
         2. We want to measure the effectiveness of IRM-SA
         3. What about efficiency? Intuitiveness? Are we going to address these or anything more?)
      ii. Mono-operation bias
         1. Only one treatment – IRM-SA.
         2. Only one case – the firefighters exercise.
         3. Is the cause construct under represented?
iii. Mono-method bias
   1. Involve different types of methods and observation for cross-checking
   2. We have expert reviews of IRM-SA designs
   3. What other metric to use?

iv. Confounding constructs and levels of constructs
   1. The presence or absence of experience in developing (or in software design) may not explain the causes of our experiment, but the difference may depend on if the students have 1, 3 or 5 years of experience → Interpret results according to profile groups constructed with questionnaire?

v. Interaction of different treatments
   1. We assign at most one treatment per subject.

vi. Interaction of testing and treatment
   1. There will be no feedback to the subjects, so they won’t have the chance to learn from their mistakes and focus more on the subjects for which they are tested.

vii. Restricted generalizability across constructs
   1. Measure more constructs than present in the hypothesis so that, e.g., we can prove that IRM-SA increases effectiveness/maintainability but at the same time does harm productivity/speed much.

b. Social threats
i. Hypothesis guessing
   1. Do not inform students about the hypothesis and the intended result, i.e., that we want to evaluate the effectiveness of IRM-SA

ii. Evaluation apprehension
   1. Since we won’t really use questionnaires for self-evaluation, it doesn’t matter.

iii. Experimenter expectancies
   1. We won’t perform interviews, so less chance to guide the subject’s answers.
   2. We will review and iterate over the experiment design and instrumentation (also involving external researchers) so that we avoid conscious/unconscious bias.

4. External validity: threats that limit our ability to generalize of our experiment to industrial practice.
   i. Interaction of selection and treatment
      1. Take care to have a subject population that is representative of the population we want to generalize to. Students can be generalized to junior software engineers [2]

   ii. Interaction of setting and treatment
1. Since design is mostly a mental process, whether it is performed in a classroom or a work place doesn’t matter. On the other hand:

2. Design in industry is usually a cooperative process performed by groups, not individuals – Do we address this? And if yes, how?

3. Design in industry usually evolves in iterations – Do we address this? And if yes, how?

4. We have a toy example (for good reasons). We can say that as a second phase we plan to experiment with (student?) pilot projects.

iii. Interaction of history and treatment

1. We will schedule the experiment for as much “neutral” days as possible.
2.8 References
