Extraction Guidelines

This document guides the extraction of relevant attributes about the analysis of experiments using a crossover design.

Throughout the guidelines, extraction rules are supplemented with examples to illustrate their application. Examples refer to specific studies via their identifier (e.g., #116). The mapping between identifiers and the study can be found in the “Studies” sheet of the data extraction file.

Extraction

During the data extraction phase, we extract relevant attributes of the data analysis of a crossover-design experiment from eligible primary studies. The extraction considers three groups of attributes:

1. **Variables**: general information about the variables investigated in the experiment.
2. **Subjects**: general information about the subjects involved in the experiment.
3. **Analysis**: main attributes of interest. Details about the inferential analysis method, the reported threats to validity associated with the crossover design, and the actions performed to address each threat.
4. **Material**: information about the accessibility of data and analysis scripts.

All attributes are extracted per experiment that meets all of the inclusion and none of the exclusion criteria, where one article can report multiple experiments.

Variables

Factors

Description: The main factors of an experiment are the variables assumed to have a causal effect on one or more response variables.

Extraction rule: Extract both the name of every main factor, all of their levels, and their measurements, as mentioned by the authors of the article.

Examples:

- In #14, the authors compare the effect of test-driven development (TDD) with traditional test-last development (TLD). The factor is the *testing approach* and its levels are *[TDD; TLD]*.
- In #116, the authors compare their own developed tool called SOCIO chatbot with an existing alternative, the Creatly web tool. The factor is the *modelling tool* and its levels are *[SOCIO chatbot; Creately web tool]*.

Response Variable(s)

Description: The response variables are the variables that are assumed to be impacted by the different levels of the main factor.

Extraction rule: For each response variable, extract both the construct as well as the measurement, as mentioned by the authors of the article. If the authors only specify the measurement without the construct that it represents, infer the construct based on context information.

Examples:

- In #9, the authors investigate the effect on *understandability* and *perceived difficulty*, respectively measured via timed actual understandability (TAU) and a questionnaire 5-point Likert scale. We report these variables and their measurement as the authors did.

Subjects

Subject number

Description: The number of human participants involved in the experiments.

Extraction rule: Extract the number of experiment subjects whose data was considered in the analysis as mentioned in the article. If the number of participants is not mentioned, record *NA*.

Examples:

- In #9, the authors report that “[a]fter the cleaning procedure, we were left with 105 valid responses”, such that the subject number is *105*.
- In #6, the authors do not mention the number of experiment participants. The subject number is recorded as *NA*.
- In #116, the authors conduct a family of three experiments with 18, 10, and 11 subjects. Each of those experiments is recorded separately with its respective subject number. Furthermore, each of these subjects consisted of a group of 3 students. Despite the actual number of involved people being higher, the number of experimental subjects is the one stated above.
- In #24, the authors report that “[t]he experiment was initially performed by 24 participants. However, 6 cases were excluded due to problems with the data collection, detected at the end of the experiment. We used the data of the remaining 18 participants.” We report a subject number of 18.

Subject Type

Description: The type of participants involved in the study.

Extraction rule: Select the appropriate categorization according to the following criteria.

Category	Criterion
Students	The participants are (university) students (including graduate students)
Student Groups	The subject of the experiment are two or more students working together
Practitioners	The participants are practitioners working in industry
Practitioner Groups	The subject of the experiment are two or more practitioners working together
Mixed Groups	The subject of the experiment are groups of at least one student and at least on practitioner
Both	The experiment involved both students and practitioners
Unknown	The authors do not specify the type of subjects participating in the experiment

Examples:

- In #6, the authors report that “[t]he subjects were undergraduate *students*”.
- In #116, the authors report that “The participants were grouped into three-member teams, where each team was considered as a subject.” The subject type is, consequently, *student groups*.
- In #9, the authors report the demographics of their study participants as the following: “59 participants were from industry (56%), 18 were professionals from academia (17%), and 28 were students (27%).” This counts as subject type *both*.
- In #77, the authors give no information about the subject type, which is noted as *unknown*.

Analysis

The analysis attributes pertain to the data **analysis of the data** obtained from conducting the crossover-design experiment. These attributes provide insight into (1) what statistical tools are used to analyze data and, more importantly, (2) whether and how threats to validity that are associated with the crossover design as detailed by Vegas et al. [1] are properly addressed.

Inferential Analysis Method

Description: The statistical method applied in order to conduct an inferential analysis of the effect of the different levels of the main factor on the response variables.

Extraction rule: Select the appropriate method according to the following criteria.

Method	Name	Criterion
NHST	Null-hypothesis significance test	One- or two-tailed test of statistically significant difference in the distribution of the response variable stratified by the levels of the main factor
GLM	Generalized Linear Model	Fixed-effects linear model, based on the maximum likelihood theory of independent observations
GLMM	Generalized Linear Mixed Model	GLM including random effects.
GEE	Generalized Estimating Equation	Parameter estimation of a GLM with a possibly unmeasured correlation based on quasi-likelihood theory with no assumption about the distribution of the response variable [2].
Other		Any other method (to be recorded in the comments)
Unknown		The authors do not state their method at all

Examples:

- In #9, the authors conduct a Mann-Whitney U test, which is a null-hypothesis significance test (coded *NHST*).
- In #50, the authors use a linear mixed effects model, coded *GLMM*.
- We are not differentiating between LMMs and GLMMs. The information relevant in this study is whether the model contains a random effect or not. Hence, all LMMs are also coded as GLMM.
- Other terms for mixed models (like “linear mixed-effects (LME) model” in #54 or “repeated measures linear mixed model” in #55) are also coded as *GLMM*.
- In #77, the authors do not really conduct any statistical test at all, but rather compare the mean values of the response variable distribution stratified by the treatment. This is coded as *Other*.
- In #86, the authors claim to conduct an analysis following some guidelines, but the actual method remains *unknown*.

Test Type

Description: The specific test type (if mentioned) conducted, e.g., when performing an NHST.

Extraction rule: If the inferential analysis method was NHST, extract the mentioned test type as reported by the authors.

Category	Criterion
Unpaired T	Parametric test for unrelated data points
Paired T	Parametric test for related data points
Mann-Whitney U	Nonparametric test for unrelated data points
Wilcoxon signed-rank	Nonparametric test for related data points
ANOVA	Parametric test for unrelated data points of two or more samples
Kruskal-Wallis	Nonparameteric test for unrelated data points of two or more samples
Other	In case none of the types fit

Threats to Validity

Description: Vegas et al. [1] mention several threats to validity - i.e., confounding factors caused by the use of a crossover design - and how to address them at analysis time. This attribute records how well the authors adhere to their recommendations.

Extraction rule: For each of the four threats to validity period, sequence, between-subject variation, carryover[1], categorize the degree of consideration based on the following criteria.

Category	Criterion
Modeled	The authors address the threat to validity by modeling the factor in the analysis (e.g., as a parameter in a GLM or GLMM).
Stratified	The authors address the threat to validity by stratifying the data by the levels of the confounding factor and conducting separate analyses.
Isolated	The authors analyze the threat to validity in isolation, i.e., conduct a statistical test with the threat variable as the only independent variable
Acknowledged	The authors do not address the threat in the analysis, but acknowledge its (unaddressed) influence in the threats to validity section.
Neglected	The authors do not address the threat to validity in the analysis, but claim it is negligible due to the employed design.
Ignored	The authors neither address nor acknowledge the threat to validity.

Examples:

- In #92, the authors analyse the data obtained from the crossover-experiment via a LMM including the treatment as well as one factor for all four threats to validity (carryover as a random effect) in the model. All threats are *modeled*.
- In #49, the authors schedule a washout period between the two experimental periods: “The execution of the experimental sessions on two different days was to have an adequate washout period between the two laboratory sessions.” The carryover threat is therefore addressed via a *washout*. We do not judge the adequacy of this approach.
- In #14, the authors conduct separate analyses where they investigate the effect of the sequence and the carryover variable on the response variable. These count as *isolated*.
- In #9, the authors acknowledge that “due to the similar task structures, learning effects are very likely.” This is an *acknowledged* threat to validity of the period variable.
- In #7, the authors explainin that they “tried to mitigate tiredness/boredom with a reduced duration of the experiment, max 25 min.” Tiredness and boredom are types

[1] We could additionally investigate the threat to validity via material, but this is confounded with either period or treatment in a 2x2 factorial design.

of period threats, meaning that the authors are aware of the potential threat to validity but did not represent it in their analysis. This counts as *neglected*.

- In #9, the authors conduct an isolated analysis of the impact of demographic factors on the response variable. The threat to validity caused by between-subject variation (i.e., individual skill) is, however, not *isolated* by this, as one could assume, as demographic factors (like skill or experience) are conceptually different from between-subject variation. The latter represents the individual skill (without any assumptions of how to model it) while the former are population-level factors. Since the authors further claim that “crossover design is fairly robust against many confounders by reducing the impact of inter-participant differences” this threat counts as *neglected* (i.e., the authors are aware of it but do not model it).
- Similarly, in #7, the authors stratify the data by *experience* and conduct separate analyses. This, again, is not the same construct as between-subject variation.
- In #6, the authors mention neither individual skill nor the carryover effect as potential threats to validity. They are coded as *ignored*.

Washout

Description: To address the threat to internal validity caused by the potential carryover effect in crossover-design experiments, some disciplines use washout periods, i.e., time between the experimental periods to diminish the influence that a previously administered treatment has on the next period.

Extraction rule: Flag the “washout” variable as true if the authors explicitly state to have included a washout period in their experimental design.

Examples:

- In #49, the authors report that “[t]he experimental sessions took place on different days in the same laboratory as the training session. The execution of the experimental sessions on two different days was to have an adequate washout period between the two laboratory sessions.” This counts as *washout*.

Material

Material refers to both the raw data generated by conducting the experiment and the scripts used to conduct the analysis. We extract these attributes to determine candidate articles where the data analysis could be reproduced.

Availability

Description: The availability represents the degree to which the material is available

Extraction rule: Select the appropriate category according to the following criteria.

Characteristic	Criteria
Archived	The material is hosted in a service satisfying all of the following criteria: ● **Immutable URL**: cannot be altered by anyone ● **Permanent**: the hosting organization has a mission to maintain artifacts for the foreseeable future ● **Accessible**: There is a DOI pointing to the real approach URL
Open Source	The material is available and has a proper license, which grants access and re-use of data, material, and source code
Reachable	The material is reachable now but is missing some aspects above to be considered Open Access.
Upon Request	The authors say the material is available upon request.
Broken	A link is given in the paper, but does not resolve or the material is no longer available.
Unavailable	Material is discussed in the paper, but no link is provided.
Private	The authors say that material exists, but it is private for some reasons (such as industry collaboration with private data, etc.).
Proprietary	The material is available but proprietary

Examples:

- In #9, the authors disclose their replication package (containing both the data and scripts) at https://zenodo.org/records/8100380. This repository has an immutable URL and is permanent (as per Zenodo's policy), accessible to all, and contains an open source license (CC4.0-BY). Hence, it counts as *archived.*
- In #64, the authors share their experimental material via Dropbox (https://www.dropbox.com/scl/fo/jp4qi7xgiamxb5z45ggjl/ACK1ECptEMJ0UN0C3a4JqV8?rlkey=w4lnd0007dlzvv29mv7oifpe3&e=2&dl=0). The files can be accessed, but neither contain an open source license not does Dropbox ensure the properties of the Archived category. Hence, it is just *reachable*.
- In #116, the authors refer to their material via https://bit.ly/34v7OTs. This link leads to a dropbox folder that no longer exists. This counts as *broken*.
- In #6, the authors mention no material at all. Their data and scripts are *unavailable*.

Location

Description: The location of the material is the URL under which the material can be accessed, according to the article.

Extraction rule: Extract the URL mentioned by the authors (if available) where the material is located.

Examples:

- In #9, the authors disclose their material at https://zenodo.org/records/8100380.
- In #64, the authors share their experimental material via Dropbox (https://www.dropbox.com/scl/fo/jp4qi7xgiamxb5z45ggjl/ACK1ECptEMJ0UN0C3a4JqV8?rlkey=w4lnd0007dlzvv29mv7oifpe3&e=2&dl=0).
- In #116, the authors refer to their material via https://bit.ly/34v7OTs.

References

[1] Vegas, S., Apa, C., & Juristo, N. (2015). Crossover designs in software engineering experiments: Benefits and perils. *IEEE Transactions on Software Engineering*, *42*(2), 120-135.

[2] Sepato, S. M. (2014). *Generalized Linear Mixed Model and Generalized Estimating Equation for Binary Longitudinal Data* (Master's thesis, University of Pretoria (South Africa)).