As mentioned in **??**, counterfactual generation is notorious for their lack of a standardised evaluation procedure. Nonetheless, we attempt to address our research questions in three experiments.

Experiment 1: Model Comparison

First, we assess the viability of **[a number of]** models. For this purpose, we sample **[10]** factuals and use the models to generate **[50]** counterfactuals. We determine the **[mean]** viability across the counterfactuals. With this experiment, we show that a model which optimizes quality criteria of counterfactuals produces better results than models, which do not. Hence, we expect the evolutionary algorithm to perform best, as it can directly optimize multiple viability criterions. In the following we list all models, we are going to compare:

RNG A Random-Search Generator , which generates random values and acts as a baseline.

CBG A Casebased-Search Generator , which samples from process instances within the training set

EVO A SBI-ES-OPC-SBM-FSR Generator , which optimizes viability using principles of evolution.

In accordance with *RQ1-H1* and *RQ1-H2* we expect the SBI-ES-OPC-SBM-FSR Generator to perform best among these baselines, when it comes to viability.

Experiment 2: Comparing with alternative Literature

The model comparison is not enough to establish the validity of our solution, as defined proposed the viability measure ourselves. Therefore, we also assess each model based on the evaluation criterions of an alternative work. More precisely, we quantify the viability of our models using the metrics employed by Hsieh et al. Hence, we measure the sparsity by computing the average Levenshtein difference and proximity using the L2-Norm. Furthermore, we compute the average intra-list-diversity and plausibility as well as the models capability of changing the prediction to a desired one.

Similar to Hsieh et al., we only focus on the *activities* that are generated by each model and its accompaniying *resource* event-attribute. For diversity and plausibility we remain close to the original evaluation protocol by Hsieh et al. as we will also treat each counterfactual trace sequence as a symbol.

Hence, a sequeunce *ABC* is treated as a completely different symbol than *ABCD*.

The goal is to show that models, which optimise viability criterions, perform better, even if viability is assessed differently as stated in *RQ2-H1* of our research question (**??**).

Experiment 3: Qualitative Assessment

For the last assessment, we follow Hsieh et al.'s procedure of assessing the models qualitatively. We use the dataset as the authors do. **[FOR XIXI: Should I use the exact same examples?]** However, as we focus on outcome prediction, we attempt to answer one of two questions:

1. *what would I have had to change to prevent the cancellation/rejection of the loan application process*

2. *what would I have had to change to get cancelled/rejected of the loan application process*

The goal is to show, that the results are viable despite not having a standardized protocol to measure their viability.