1 Take-Home Message

We wanted to design a single-case experiment to test whether classical conditioning can influence pain thresholds. We settled on a customized AB phase design for which we ran pilot tests and a power study. The results indicated that this single-case design can be used to test our hypothesis.

2 Abstract

Single-case experiments are increasingly popular in the behavioral sciences. Due to their flexibility, single-case designs can be customized to test a variety of experimental hypotheses. We were interested in using a single-case experimental approach to test whether pain thresholds can be influenced by Pavlovian classical conditioning. Following the example of earlier studies into this topic, we planned to measure whether participants would more frequently report specific electrocutaneous stimuli as painful when they were presented with specific vibrotactile stimuli that had previously been associated with painful electrocutaneous stimuli. First, we decided on a mean difference effect size measure derived from the Sensation and Pain Rating Scale ratings for the electrocutaneous stimuli provided by the participants. Next, we discussed several possible single-case designs and evaluated their benefits and shortcomings. Then, we ran pilot tests with a few participants based on the possible single-case designs. We also conducted a simulation study to estimate the power of a randomization test to test our hypothesis using different values for effect size, number of participants, and number of measurements. Finally, we decided on a sequentially replicated AB phase design with 30 participants based on the results from the pilot tests and the power study. We plan to implement this single-case design in a future experiment to test our hypothesis.

Keywords : single-case, classical conditioning, pain, randomization test, statistical power.

3 Purpose

In this manuscript, our purpose is to describe how we used a trial and error approach to design a customized single-case experiment (SCE) to test the effect of classical conditioning on pain thresholds. We hope that this information helps readers who plan to implement an SCE in better understanding the steps involved and possible challenges. We wish to focus on using the flexible nature of SCEs to adapt to the requirements of the experiment. We also wish to highlight the necessity of exploring various available statistical methods and statistical power requirements for these methods prior to starting the experiment. Finally, we demonstrate how SCEs can serve as a viable avenue for pilot-testing new therapies and treatments at a small scale before implementing larger scale studies.

4 Introduction

4.1 Classical Conditioning

Persistent pain—pain that is still felt after bodily tissue has healed—is a major healthcare problem that is poorly understood and, consequently, difficult to treat effectively . While there are key models that go some way towards explaining how pain can persist without active tissue damage (**bib1**; **bib2**; **bib3**; **bib4**) , certain pain presentations remain unexplained. The imprecision hypothesis was proposed to address these unexplained presentations, and is founded on the idea that pain can be modulated by pain-associated cues—specifically, via "classical conditioning" (**bib5**) .

Our study design builds on two previous experiments (**bib6**; **bib7**) , which followed a Pavlovian or "classical" conditioning design (**bib8**) to test whether neutral but pain-associated cues can bias a participant's decision about whether a stimulus is painful or non-painful. In the differential classical conditioning design used here, one neutral cue is paired with a painful electrocutaneous stimulus, while another neutral cue is paired with a non-painful electrocutaneous stimulus. It is expected that participants form associations between each cue and painfulness, such that subsequent perception of an ambiguous electrical stimulus is modified, dependent on the simultaneous presentation of one of the neutral cues.

We set out to design an SCE to test this hypothesis. Due to the before (conditioning) and after (conditioning) structure of multiple observations required for each participant, SCE designs are perfectly suited for testing this hypothesis.

4.2 Single-Case Experiments

SCEs are experiments in which the effect of manipulating an independent variable is observed in a single entity (**bib9**; **bib10**) . Variables of interest in the entity, which is often a single participant, are measured repeatedly over a period of time with the purpose of establishing a causal relationship between the independent variable, commonly referred to as the treatment condition, and the observed entity. SCEs are increasingly popular in behavioral and educational sciences due to their flexibility, low cost, and focus on effects of the intervention in individual participants (**bib11**) . SCEs may be preferable to group-based designs, where the existence or magnitude of the effect *within* an individual is more relevant than the average effect *across* a group of individuals. The lower cost and flexibility of SCEs also make them ideal for studying new theories and interventions, as is the case in this study (**bib12**) .

In SCEs, assignment of the treatment conditions to measurement occasions can be randomized to strengthen internal validity (**bib13**; **bib14**) . Additionally, replication using multiple participants strengthens external validity and generalizability of the results (**bib13**; **bib15**) . In our study design, we incorporated both randomization and replication to enhance internal and external validity of the experiment.

4.3 Advantages of SCEs for Classical Conditioning

SCEs are well suited to research questions about within-individual processes. Although group designs are historically respected, they are typically analyzed in a way that aggregates data from all individuals to estimate an effect at the group level—an effect that may be non-existent in any of the individuals within the group (**bib16**) . This aggregation approach obscures the true, within-individual effect, and the between-individual variability of that effect (which, itself, is typically worthy of attention; (**bib17**) . Of course, it is possible to consider within-individual changes in some group designs, but an SCE offers a superior approach to achieving a sufficiently powered ($> 80\%$) examination of within-subject effects (**bib18**; **bib19**) . Whereas traditional group-based designs generate knowledge on the population level only, SCEs generate knowledge at the level of the individual. Finally, findings from group-based designs cannot be generalized towards individuals, while SCEs can be aggregated for insights on populations.

4.4 Analyzing SCEs

SCEs are often difficult to analyze due to the presence of serial correlation in observed data and nonadherence to distributional assumptions (**bib20**; **bib21**; **bib22**; **bib23**) . Traditionally, visual analysis was the preferred method for analyzing SCE data, but researchers now recommend quantitative statistical analysis to complement visual analysis (**bib24**; **bib25**; **bib26**; **bib27**; **bib28**) . Specifically, for experiments involving user-reported ordinal data, such as the current experiment, nonparametric statistical methods may be best suited for analysis, as the observed data may not adhere to distributional assumptions required for popular parametric methods.

Randomization tests (RTs) are nonparametric hypothesis tests recommended for analyzing SCEs (**bib14**; **bib22**; **bib11**) . RTs derive their validity from the random assignment of treatment conditions to experimental units, which in the case of SCEs, are measurement occasions (**bib29**) . RTs do not require any distributional assumptions for the observed data. Additionally, RTs are immensely flexible, and can be adapted for any SCE design, randomization scheme, and effect size measure as the test statistic (**bib30**; **bib31**; **bib11**) . In our study design, we implemented a customized RT using a mean difference effect size measure as the test statistic.

4.5 Statistical Power of SCE RTs

Power analysis is an essential part of statistical hypothesis testing, particularly for determining sample size (**bib32**) . Several guidelines recommend a power analysis or at least a methodical description of how sample size was determined (**bib33**; **bib34**) . In the context of SCEs, power analysis can be used to determine the number of measurements necessary, and for studies with replication, the number of participants.

For RTs, the randomization distribution is calculated empirically, and hence statistical power can only be estimated using computer-intensive simulations (**bib35**; **bib18**) . Several studies have estimated the power of SCE RTs for various design conditions (**bib19**; **bib18**; **bib36**; **bib37**; **bib38**; **bib39**) . Although the results from these studies can be used as a reference, they are true only for the design parameters and simulated data distributions considered. Since we acquired unusual ordinal observed data, power ideally needed to be estimated using simulated data of the same type as this study would generate.

5 Methods

5.1 Experimental Setup

In this experiment, participants receive two different types of stimuli: a vibrotactile stimulation as the conditioned stimulus (CS) and an electrocutaneous stimulation as the unconditioned stimulus (US). At first, one neutral stimulus (a vibrotactile stimulus, denoted as CS+) is paired with a painful stimulus (a high-intensity electrocutaneous stimulus, denoted as US), whereas a different neutral stimulus (a vibrotactile stimulus, denoted as CS-, applied to a different location) is paired with a non-painful stimulus (a low-intensity electrocutaneous stimulus, denoted as US). This procedure is expected to form an association between the CS+ and painfulness, in contrast to the CS- and non-painfulness. Later, each CS is paired with a "test stimulus" (denoted as US), an electrocutaneous stimulus of an intensity calibrated such that there is a 50% chance that the participant judges it to be painful or non-painful at baseline (i.e., before the pairing). Participants are then requested to judge each trial on a scale that distinguishes painful from non-painful events (**bib40**) . The trials of interest are the CS/US pairs, as the primary hypothesis is that, due to the pairings learned early in the experiment, the participant will come to judge the US stimuli to be painful more often when they are paired with the CS+ (previously paired with a painful stimulus) than when they are paired with the CS- (previously paired with a non-painful stimulus).

5.2 Laboratory Setup

The participants receive two types of stimuli at the same time, replicating the procedure in Traxler et al. (**bib7**) . Stimulus onset and timing are controlled using Affect 4.0 (**bib41**) . A vibrotactile stimulus is delivered to the skin using tactors manufactured by Dancer Design taped to the participant's skin (**bib42**) . This stimulus is of fixed duration and a clearly perceptible intensity. Three tactors are used in the arrangement as the source of a CS (see (**bibFigure 1**)). The SIDE tactor is allocated to CS , whereas the ABOVE and BELOW tactors are assigned to CS+ and CS- in a counterbalanced way across participants.

An electrocutaneous stimulation is delivered by passing a current across two surface electrodes located at the midpoint between the tactors to serve as the US

Figure 1: *Arrangement of Tactors and Electrodes on the Back (Adapted From Traxler et al.* (**bib7**) *) .*

(Figure 1). The current is delivered using a DS7A constant current stimulator (**bib43**) . The stimulation intensity is calibrated individually. The number of pulses delivered on each stimulation occasion is varied to provide a US that is usually painful, a US that is usually non-painful, and a US that is calibrated to lie close to the pain threshold, the boundary between non-painful and painful.

5.3 Tactor and Electrode Preparation

On arrival, participants are seated straddling a chair in front of a desk with a computer monitor, mouse, and keyboard. Participants are asked to bend backwards to identify the point at which the greatest bend is seen. A point in the upper lumbar region is marked on the back, 2cm to the left of the spine, where the electrodes are placed such that the mark lies exactly between them. Three other points—4cm above, below, and to the left side of the electrodes—are marked, and the tactors are taped in place such that the closest border of each tactor lies at a tactor mark (Figure 1). Calibration is performed by the procedure described in Traxler et al. (**bib7**) , with the CS paired with each electrocutaneous stimulus. The CS is used to provide vibratory stimulation consistent with the CS+/CS- stimuli that are later presented in the experimental trials to ensure that any modulation of pain by the vibration itself is consistent across calibration and experimental trials. Ratings of CS trials from the experimental phase are not relevant to the research question and are therefore not analyzed.

5.4 Experimental Trials

Based on the experimental setup discussed previously, and given the pairing of CSs and USs, the experiment for a participant can theoretically consist of five different types of trials:

1. Training trials: Each training trial consists of one CS, with each of the three CS types used in different trials. There is no electrocutaneous stimulation (i.e., US) presented. These trials are conducted to familiarize the participants with the locations of the CS tactors.

2. Baseline trials: The baseline trials consist of two combinations of stimulations: a CS+ with a simultaneous US and a CS- with a simultaneous US . These trials are used to record baseline pain ratings of participants before they are conditioned to associate the painful US with a particular CS.

3. Acquisition or learning trials: These trials consist of two combinations of stimulations: a CS+ with an US and a CS- with an US , and are meant to condition participants to associate painful US to CS+ and vice versa.

4. Test trials: These trials are the same as the two baseline trials: a CS+ with an US and a CS- with an US , and are meant to record pain ratings of participants to test whether they have associated the painful US with CS+ and vice versa.

5. Irrelevant trials: These trials consist of a CS combined with an US stimuli. These trials are noninformative and are included to satisfy the assumptions of the analytical approach; they are included only to achieve consistency in the time gaps between baseline or test trials.

5.5 Observed Variables

The participants are requested to provide two reports on each trial. First, they rate the stimulation event on the Sensation and Pain Rating Scale (**bib44**; **bib40**) . The SPARS is anchored at -50 (no sensation), 0 (the exact point at which the feeling transitions to pain), and 50 (worst pain imaginable). The two distinct ranges for non-painful (-50 to -1) and painful (1 to 50) are clearly marked and explained to the participants. Participants are explicitly advised to make an initial decision about whether a trial was non-painful or painful before assigning a rating from the appropriate side of the scale, without selecting 0 on SPARS.

Second, the participant indicates the location at which they feel the vibrotactile stimulus from the three options "ABOVE", "BELOW", and "SIDE". This is only used to confirm that the participants are identifying the location of the stimuli correctly. Given that discrimination of the vibrotactile stimuli is necessary for differential learning, participants who fail to identify the location of the stimulus correctly in at least 75% of the trials are excluded and replaced by additional participants.

5.6 Effect Size Measure and Test Statistic

We are interested in calculating whether the participants judged trials of CS+/ US and CS-/ US as painful or non-painful differently for baseline and test trials. For this purpose, we first convert the SPARS ratings to a binary variable: 0 for non-painful (-50 to -1 on SPARS) and 1 for painful (1 to 50 on SPARS). A rating of 0 on SPARS can neither be classified as painful nor as non-painful, therefore these are considered indeterminate and were marked as missing. Then, we pair two trials, a CS+/ US and a CS-/ US and calculate the difference between the corresponding binary variables. The resulting difference indicates whether the participant is rating CS+/ US and CS-/ US differently. The difference can result in three different values: 1 when the CS+/ US trial is rated as painful while the corresponding CS-/ US trial is rated as non-painful; -1 when the CS+/ US trial is rated as non-painful while the corresponding CS-/ US trial is rated as painful; and finally, 0 when both CS+/ US and CS-/ US trials are rated as painful or both are rated as non-painful. This difference value, calculated from a pair of trials, constitutes one measurement in our SCE. A simple mean

difference (MD) effect size measure is calculated as the difference between the means of the difference values derived from the test trial pairs and the difference values derived from the baseline trial pairs. Due to the flexibility of RTs, this effect size measure can also be used as the test statistic (**bib31**) .

5.7 Single-Case Design

In this section, we describe our considerations regarding which SCE design to use in our experiment. We discuss several initial design possibilities that were chosen by trial and error and conclude with a final design.

5.7.1 Initial Design Options

Randomized Block Design. Since each of our measurements requires a pair of trials, we immediately considered a randomized block design (RBD) for the experiment. In an RBD SCE, similar to an RBD in traditional group designs, the measurement occasions are divided into small blocks, with each block containing administrations of all treatment conditions (**bib45**) . For our study, a possible block consists of the two baseline (or test) trials, a CS+/ US and a CS-/ US , in random order. Hence, the experiment for a participant can consist of several blocks of baseline trial pairs, followed by a period of acquisition trials to condition the participants, followed by several blocks of test trial pairs. However, this design is immediately rejected, as an RBD SCE requires all treatment conditions to be applied inside a block. Since the effects of conditioning are acquired over a period of time and are expected to carry over to future trials, true alternation of treatment conditions is not possible in this experiment.

AB Phase Design with In-Phase Acquisition. When we rejected the RBD, we realized that the experiment should consist of periods of baseline trial pairs and periods of test trial pairs. The simplest method to achieve this is using an AB phase design, with an A phase as the baseline condition and a subsequent B phase as the test condition (**bib9**) . In favor of internal validity, the phase transition is preferably randomized. Each phase consists of several blocks of trials, each of which yields one measurement. Each measurement block in the baseline phase consists of k irrelevant trial pairs (two CS / US) and a baseline trial pair (a CS+/ US and a CS-/ US in random order). Similarly, each measurement block in the test phase consists of k acquisition trial pairs (a CS+/ US and a CS-/ US in random order) followed by a test trial pair (a CS+/ US and a CS-/ US in random order). The number of irrelevant or acquisition trial pairs in each block, or k , is set based on the required level of acquisition. A lower k would mean we would need a lower number of trials to achieve a certain number of measurements, but would also typically result in weaker acquisition of associations, and vice versa for a higher k . Unfortunately, this design has low acquisition of the associations as there is no learning period. Additionally, RTs for AB phase designs require a large number of measurements to achieve sufficient statistical power (**bib46**; **bib18**) . As a result, the experiment for a participant would need a very large number of trials over a significantly long

time to ensure both high acquisition and enough measurements. As a result, this design is rejected in favor of a slightly modified AB phase design. A similar ABAB phase design is also rejected due to the potentially large number of trials required and possibility of carry-over effects between phases.

5.7.2 Final Design

The final design combines elements from both previous designs: an AB phase design with both out-of-phase and in-phase learning. The baseline phase for this design contains several pairs of baseline trial pairs (a CS+/ US and a CS-/ US in random order), followed by a period of learning with only acquisition trial pairs (a CS+/ US and a CS-/ US in random order). Finally, the test phase includes test trial pairs (a CS+/ US and a CS-/ US in random order) with a few acquisition trial pairs in between at regular intervals to reinforce the associations and hence prevent extinction. The time of onset of the test phase is randomized. The experiment starts with a few training trials before the baseline phase.

The number of trials and the number of participants are to be decided based on a power study and some pilot experiments. However, the guidelines for phase designs by Kratochwill et al. (**bib26**) recommend at least five measurements in both baseline and test phases. Since we are also curious about any increase or decrease in acquisition over time, we decide to include at least 15 measurements in the test phase.

Due to the use of specialized equipment and individual calibration, multiple participants cannot be tested at the same time. Hence, simultaneous replication using a multiple baseline design (MBD), which is able to control for environmental confounding factors, is not possible (**bib9**; **bib10**) . Instead, the experiment is to be run as a sequentially replicated AB phase design experiment.

5.8 Randomization Scheme and RT

The design is randomized using intervention start point randomization for each participant (**bib14**) . In this scheme, the total number of measurements and the minimum number of measurements in each phase is decided beforehand, and the intervention can start at any time point that satisfies these restrictions. Therefore, if the number of measurements is N , and the minimum number of measurements in A and B phase are a and b respectively, the number of possible randomizations is $r = N - a - b + 1$. For P participants with the same randomization scheme, the total number of randomizations for the replicated experiment is r .

We conduct a single RT for all participants combined based on these r randomizations. The null hypothesis for this RT is that no difference exists between the baseline phase measurements and test phase measurements for all participants. As discussed previously, the measurements represent whether participants rate CS+/ US trials as painful more frequently than CS-/ US trials. Since the Pavlovian conditioning procedure is aimed at participants perceiving

the CS+/ US as more painful, the RT is one-sided with the alternate hypothesis being that the test phase measurements are higher than the baseline phase measurements for at least one participant. This RT can be conducted with the average (across participants) of the MD effect size measure discussed previously as the combined test statistic.

We use the Shiny SCDA (Single-Case Data Analysis) web app for SCEs to randomize and analyze the experiment (**bib47**) . Whereas this web app does not include an option for simultaneously replicated AB phase design, it does include an MBD option. The MBD option in Shiny SCDA uses the Koehler-Levin regulated randomization procedure (**bib48**) , which, when using an identical set of possible start points for all participants, is analogous to our randomization scheme. Hence, we can both randomly select test phase start points and run the RT in Shiny SCDA. Additionally, we can plot the observed data for visual analysis in Shiny SCDA.

5.9 Pilot Tests

We ran several pilot tests to estimate possible effect sizes and to test possible design choices. We first tested nine participants under slight variations of our initial design choice. Later, after a simulated power study and adjustments to the design, we tested another four participants. We analyzed these four tests using visual analysis and an RT in Shiny SCDA. These four tests were, however, not randomized, and the test phase for all four participants started at the tenth measurement occasion. However, for the purposes of demonstration, we ignored the assumption of randomization. We will discuss the results from these pilot tests, and the final four tests in particular in the Results section.

There were two important takeaways from the first nine pilot tests that were very useful for the design of the power study. First, the observed MD effect sizes were extremely small. The average effect size from the first nine tests was slightly negative at -0.037, with a maximum of 0.143 and a minimum of -0.400. Second, the pilot tests also revealed difficulties in running the experiment for more than 30-35 minutes for a participant. Considering each trial took around 15 seconds, this gave us a maximum of around 140 trials.

5.10 Power Study

To ensure sufficient statistical power for our RT, we needed to first estimate power for different values of number of measurements (N) and number of participants (P), and then choose sufficiently high values of N and P for the experiment.

Ideally, we would want to select the maximum N possible within the limits of how long an experiment can be reasonably run. This strategy presents us with two advantages. First, increasing N should result in reduction in the variability of the estimate of effect size. Second, maximizing N should allow us to achieve sufficient power with lower P , which directly equates to lower cost for the experiment.

We used the Monte Carlo method used by Ferron and Onghena (**bib19**) to estimate power. In this method, several datasets are simulated under a set of simulation conditions. The proportion of these simulated datasets in which the RT leads to a rejection of the null hypothesis gives an estimate of statistical power for the given set of simulation conditions. A simulation study of such complexity is both difficult to program and computationally intensive to execute. Fortunately, we were able to modify and repurpose R code used by De, Michiels, Tanious, et al. (**bib39**) for this study.

5.10.1 Simulation Conditions

The following three simulation conditions were varied for this power study:

1. Effect size: Based on the low effect sizes observed in the pilot tests, we simulated MD effect sizes of 0.1, 0.2, 0.3, 0.4, and 0.5 for the simulated observed data.

2. Number of measurements: We decided on a minimum of five measurements in the baseline phase and a minimum of 15 measurements in the test phase. As a result, the experiment needs more than 20 measurements. However, from the pilot tests it was evident that an experiment with more than 140 trials (70 trial pairs) was not feasible. Considering around 30 trials are required for the training and learning periods, and a few more acquisition trails for reinforcement during test phase, it was difficult to have more than 40 measurements (80 trials). Hence, we used 25, 30, 35, and 40 measurements for simulated data.

3. Number of participants: Since the pilot tests revealed a small effect size, and the number of measurements is also limited, to ensure high power we considered a large range of participant count at 10, 20, 30, 40, and 50.

5.10.2 Other Simulation Parameters

For the power study, we wanted to simulate observed values that were similar to the experiment. For the baseline, we assumed a symmetric distribution around 0 for the SPARS rating. Therefore, the painful/non-painful ratings were expected to be 0s and 1s equally for both CS+/ US and CS-/ US trials. Hence, an observed value corresponding to a trial pair in the baseline was expected to be 1 with 25% probability, -1 with 25% probability, and 0 with 50% probability. We simulated observed values as 1, -1, and 0 with these probabilities. For the test phase, we increased the probability of 1 by half the selected effect size and decreased the probability of -1 by half the selected effect size. This resulted in an expected MD effect size equal to the selected effect size.

Since the number of randomizations for the RT was huge, we simulated Monte Carlo RTs with 1000 randomizations (**bib49**) . For estimating power, we simulated 10000 datasets for each set of simulation conditions. Finally, we used a 5% level of significance for the simulated RTs.

Table 1: *Estimated P ower S imulated U sing D ifferent V alues for E ffect Si ze, N umber of M easurements, and N umber of P articipants.*

No. of measurements	No. of participants	Effect size				
0.1	0.2	0.3	0.4	0.5		
25	10	10.3	19.0	32.2	48.6	66.1
	20	13.4	29.3	51.4	74.4	90.0
	30	16.4	38.8	66.1	87.9	97.7
	40	18.9	47.5	77.4	94.8	99.5
	50	22.2	54.7	85.1	97.9	99.9
30	10	13.0	25.9	45.0	66.1	83.9
	20	17.9	42.0	70.6	90.8	98.5
	30	21.9	54.9	86.0	97.9	99.9
	40	26.4	65.2	92.7	99.6	100.0
	50	30.0	73.4	96.6	99.9	100.0
35	10	14.3	31.4	55.0	77.4	92.7
	20	21.0	51.4	81.0	96.6	99.7
	30	27.4	65.5	93.3	99.7	100.0
	40	32.8	77.1	97.5	99.9	100.0
	50	37.4	84.7	99.1	100.0	100.0
40	10	15.4	36.0	63.9	84.9	96.3
	20	23.6	58.1	88.4	98.6	100.0
	30	30.7	74.0	96.8	99.9	100.0
	40	37.0	84.6	99.2	100.0	100.0
	50	42.8	90.4	99.8	100.0	100.0

The simulations were run on supercomputer nodes at the Flemish Supercomputer Center (Leuven, Belgium). This allowed testing more simulation conditions and achieve high accuracy simulating a large number of datasets for each simulation condition; however, the simulations can also be run at a smaller scale on a personal computer.

6 Results

6.1 Power Analysis

The results from the power study (1) revealed that due to the relatively small effect sizes and restricted number of measurements, the number of participants need to be high to achieve 80% power. If we restrict the number of measurements to 30, which allows sufficient acquisition trials for learning and reinforcement within 30 minutes, 30 participants result in sufficient power even with a moderate effect size.

Based on these results, we decided to run the final set of pilot tests with 30 measurements. We decided on six training trials (two trials for each type of CS) before the baseline phase, 24 acquisition trials between the baseline and test

Figure 2: *Final Design With 30 Measurements.*

Figure 3: *Plot of Observed Scores Obtained From the Initial Nine Participants in the Pilot Tests.*

phases, and additionally one acquisition trial pair for reinforcement after two test trial pairs during the test phase. With a minimum of five measurements in the baseline phase, and a minimum of 15 measurements in the test phase, this design allows for 11 possible randomizations. Depending on the number of measurements in the test phase which will vary based on the test phase start point selected, the experiment could theoretically consist of 104 to 114 trials. This is the design we intend to follow in the final experiment with 30 participants ((**bibFigure 2**)).

6.2 Pilot Tests

As mentioned previously, the initial nine pilot tests resulted in extremely small (and a few negative) effect sizes. The average MD effect size was -0.037, with a maximum of 0.143 and a minimum of -0.400. The low effect sizes indicated that the final design would need a large number of measurements and participants to achieve sufficient power in the RT. The first seven of these pilot tests consisted of 10-20 measurements each, which seemed too few. On the other hand, the last two pilot tests out of these consisted of 238 trials each. The feedback from both the experimenters and participants was that these tests were too long. These results influenced the decision to limit the number of trails to 140, remove irrelevant trial pairs from the baseline phase, and lower the number of acquisition trail pairs in the treatment phase.

The later four pilot tests were run using the design parameters we decided on after the power study. The average MD effect size for these participants was 0.057, with a maximum of 0.476 and a minimum of -0.250. Visual analysis ((**bibFigure 4**)) revealed that the first, third, and fourth participant did not seem to show any sign of conditioning. However, the second participant seemed to show significant conditioning effects. This is also confirmed by the MD effect size for the second participant (0.476). Finally, a Monte Carlo RT with 1000 randomizations resulted in a p -value of 0.112. Hence, the null hypothesis of the RT could not be rejected at a 5% level of significance.

Figure 4: *Plot of Observed Score s O btained F rom the F inal F our P articipants in the P ilot T ests.*

7 Discussion

The power analysis and pilot test results confirm that the SCE design developed for this study can be effectively used to test the effect of classical conditioning on pain thresholds. The power study provides strong evidence that this design results in sufficient power if the number of measurements and participants are chosen correctly. The results from the final four pilot tests were encouraging. Even though the RT lacked power due to the small number of participants, the p -value was low. The visual analysis also seemed to suggest a large effect for at least one participant.

The final study using this protocol was not conducted immediately, due to lack of resources at the time. However, we hope to conduct this study as soon as the opportunity arises. Meanwhile, it seemed that the discussions regarding the SCE design and preparations for the study might be useful for other researchers developing similar protocols. Therefore, we decided to prepare this manuscript.

The final design suffers from a few limitations. The first limitation is the observed variable defined by us, which only yields values of -1, 0, and 1. Unfortunately, this restricts variation in observed data and can cause duplicates in the randomization distribution, which can affect power. An alternative is to use the difference between SPARS ratings from CS+/ US and CS-/ US trial pairs. However, this would require a slightly different hypothesis, which would not clarify whether classical conditioning can affect pain thresholds specifically.

The second limitation is due to the attributes of the AB phase design. RTs using randomization of intervention start points in AB phase designs are known to lack power at lower sample sizes (**bib46**) . AB phase designs do not satisfy the guidelines set by Kratochwill et al. (**bib26**) without multiple replications. Since we are not sure how quickly the effect of our conditioning is reversed, we cannot use phase designs with more phase changes, such as the ABAB phase design. Due to the requirement of specialized equipment, we cannot use an MBD. Critically, we cannot increase the number of measurements due to time constraints. Therefore, we have to rely on sequential replications for both validity and statistical power.

The power study for this design also presents certain limitations. We used only one possible distribution of the SPARS ratings to simulate our observed data. While the assumption of a symmetric distribution around 0 is not unreasonable, with more pilot data, it might be possible to simulate using a distribution that resembles the observed data. We also did not account for any variability in effect size across participants. Instead, we simulated an equal effect size for all participants. However, we believe these are reasonable assumptions given the scope of our study.

Finally, we conducted Monte Carlo RTs for both the power study and the pilot data. As discussed previously, the number of possible randomizations for this design is r , where r denotes the number of possible randomizations for one participant, and P denotes the number of participants. Even for the smaller scale of the pilot data, computing 11 possible randomizations would have been extremely costly. Monte Carlo RTs present a simple alternative to this

computation cost while maintaining sufficient statistical power (**bib49**; **bib50**) . Hence, we intend to conduct a Monte Carlo RT in the final experiment.

As introduced earlier, this design presents certain advantages over traditional group study designs. Single-case designs are a better match for studying within-person changes, because they allow detailed insight into within-individual processes rather than assuming that the response of all individuals in a group is consistent. Additionally, these designs typically require a smaller number of participants. This allows for a lower overall cost and is particularly important for pain studies because it can be difficult to recruit participants.

8 Conclusion

In this manuscript, we described how we used trial and error to design an SCE testing whether classical conditioning can affect pain thresholds. We used a sequentially replicated AB phase design and conducted a simulated power study to determine sample size. We decided on 30 participants and 30 measurements per participant. Finally, we ran some pilot tests using our design. While the results from the pilot tests were inconclusive, they were sufficiently encouraging that we plan to conduct the full study in the near future.

9 Funding

This research is supported by the Asthenes long-term structural funding - Methusalem grant (nr. METH/15/011) by the Flemish Government, Belgium. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Centre), funded by the Research Foundation - Flanders (FWO) and the Flemish Government. Research collaboration supported by the IASP Developing Countries Collaborative Research Grant. VJM is supported by the Fogarty International Center of the National Institutes of Health (award K43TW011442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

10 Data Availability

The data and R code used in this study are openly available on Open Science Foundation at https://osf.io/d9mfn/?view_only=6b9bd4580dbf42b4952e551972f48c68 .

11 Appendix

Table 2: *Observed Scores From the Initial Nine Pilot Tests.*

P1	V1	P2	V2	P3	V3	P4	V4	P5	V5	P6	V6	P7	V7	P8	V8	P9	V9
A	0	A	1	A	1	A	0	A	NA	A	-1	A	0	A	1	A	0
A	1	A	0	A	0	A	0	A	0	A	0	A	0	A	1	A	0
A	0	A	0	A	0	A	1	A	0	A	-1	A	0	A	1	A	0
A	0	A	0	A	0	A	1	A	0	A	1	A	0	A	1	A	-1
A	1	A	0	A	0	A	-1	A	0	A	0	A	0	A	0	A	0
B	1	B	0	A	0	A	0	A	0	A	0	A	0	A	-1	A	0
B	1	B	0	A	NA	A	0	A	0	A	0	A	0	A	0	A	0
B	0	B	0	A	1	A	0	A	0	B	0	B	0	A	0	A	1
B	0	B	-1	A	1	A	1	A	0	B	0	B	0	A	0	A	0
B	0	B	0	A	0	A	-1	A	0	B	0	B	0	B	1	B	0
				B	0	B	1	B	-1	B	0	B	0	B	0	B	0
				B	NA	B	0	B	0	B	0	B	0	B	0	B	0
				B	NA	B	0	B	0	B	0	B	0	B	0	B	0
				B	0	B	0	B	0	B	0	B	0	B	0	B	0
				B	0	B	0	B	0	B	0	B	0	B	1	B	-1
				B	0	B	-1	B	0					B	0	B	1
				B	0	B	0	B	0					B	0	B	0
				B	0	B	1	B	0					B	1	B	0
				B	NA	B	0	B	0					B	1	B	-1
				B	1	B	1	B	0					B	1	B	0
														B	0	B	1
														B	0	B	1
														B	1	B	0
														B	1	B	0
														B	0	B	0
														B	1	B	0
														B	0	B	0
														B	0	B	0
														B	0	B	0

Table 3: *Observed Scores From the Final Four Pilot Tests Formatted for Shiny SCDA.*

P10	V10	P11	V11	P12	V12	P13	V13
A	1	A	0	A	0	A	0
A	NA	A	0	A	0	A	0
A	0	A	0	A	0	A	0
A	1	A	0	A	0	A	0
A	0	A	0	A	0	A	0
A	0	A	0	A	0	A	0
A	0	A	0	A	0	A	0
A	1	A	0	A	0	A	0
A	-1	A	0	A	0	A	0
B	NA	B	1	B	0	B	0
B	0	B	0	B	0	B	0
B	NA	B	1	B	0	B	0
B	NA	B	0	B	0	B	0
B	NA	B	1	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	1	B	0	B	0
B	0	B	1	B	0	B	0
B	0	B	1	B	0	B	0
B	0	B	1	B	0	B	0
B	0	B	1	B	0	B	0
B	0	B	1	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	1	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	0	B	0	B	0
B	0	B	0	B	0	B	0