Appendix A: The Propositional Model of Conditioning

Mitchell et al. (2009) have proposed a new model of conditioning that sees the formation of propositions as the main mechanism rather than the formation of associations. Propositions are *"qualified mental links, that is, links that specify how two events are related"* (p. 186). An example of this can be a belief that a specific sound precedes an electric shock.

In the context of the placebo effect, the propositional model has not been generally acknowledged by researchers other than De Houwer (2018), one of its founders. According to his view, this model could account for some of the placebo phenomena, such as the role of expectations and prior experience influencing the placebo effect. This advantage is mostly related to one of the model's key assumptions that even though propositional beliefs are accessible to conscious processes, the mechanism of forming propositions is automatic (De Houwer, 2009). This is in accordance with research showing that conditioning placebo effects can be modulated by means of changing expectations (described in the Expectation Models subsection).

The model itself has started an extensive conceptual debate (see the open peer commentaries to Mitchell et al., 2009). For example, Dickinson (2009) has pointed out that the association formation models are compatible with nonautomatic processes as well and that unlike these models, Hower has not sufficiently accounted for the relation between automatic and non-automatic processes. Similarly, a part of the debate has been focusing on the differences between animal and human conditioning with arguments both supporting (Chater, 2009) and questioning (Castro & Wasserman, 2009) the propositional model.

Appendix B: The Role of Culture

Various psychological models incorporate at least partially the idea that the placebo effect can be culturally dependent. They do not reason that the placebo effect would differ among nations in and of itself, but rather that different cues are more likely to trigger the placebo effect in different cultures (Colloca & Miller, 2011; Wickramasekera, 1980).

One such phenomenon might potentially be the effect of placebo colours. In general, it has been suggested that red, orange and yellow placebo pills produce more stimulating effects, while blue and green are more related to tranquilising effects (de Craen et al., 1996). Colour perception might be a culturally dependent symbol. A study examining colour related emotions found that there was little agreement among British and Chinese ratings of colours on the tense-relaxed and like-dislike scales (Ou et al., 2004). While blue seems to be the most preferred colour across different cultures and there is a similar cross-cultural pattern for colour clustering, there are significant cultural discordances in both the proximity clustering and in preferences for colour pairings (Madden et al., 2000). In accordance with previous research, Wan et al. (2015) have found that the colour of pills influenced the participants' perception and response expectancies. Moreover, expectations associated with different colours and shapes of pills slightly varied among Chinese, Colombian, and North American participants. For example, while red pills tended to be perceived as the most alerting and white pills were perceived as the most effective for the treatment for headache across all three of the cultures, only Chinese participants expected red and blue tablets to be harder to swallow.

Moerman (2000) has analysed data from RCTs on medication for ulcer disease, hypertension, and generalised anxiety disorder. The healing rates in placebo control groups varied considerably across the 32 countries. For example, while the average healing rate in control groups for ulcers was 36 %, this number was almost doubled in Germany (59 %) and only 22 % in Denmark and the Netherlands. On the other hand, the data for hypertension showed the opposite trend with Germany having the smallest improvement in control groups. It is not known if these differences are related to genuine placebo effects, or if they are more related to other factors such as participant selection. That being said, it is possible that cultural factors affecting expectations, such as general trust in the healthcare system (and the general attitudes towards various conditions), might lead to placebo effects of different effect sizes. This is in accordance with research showing that trust in a healthcare provider is associated with better treatment outcomes (Murray & McCrone, 2015).

Moreover, culturally specific beliefs have been suggested to influence the proneness to disease. Philips et al. (1993) have examined death records of 28 169 Chinese-Americans compared to the death records of 412 632 controls matched in their age of death, cause of death, and other factors. Chinese-Americans died 1.3 – 4.9 years sooner than their American controls when there was a match between a disease (the cause of death) and a date of birth predicting proneness to the disease according to Chinese astrology. The relationship was stronger among Chinese-Americans whose families refused necropsy to be performed (an indirect measure of adherence to the traditional Chinese culture). Moreover, the relation was more pronounced in acute diseases rather than chronic ones, therefore suggesting that factors such as lower adherence to treatment as a result of negative beliefs were not likely to be the main cause.

Although there is a considerable body of research on the role of culture in health, studies examining placebo effects in relation to culture are scarce and the specific factors remain unknown.

Appendix C: Bias in Placebo Research

Challenges	Characterization	Mechanism	Likely impact
Selection bias	Selection of patients or experimental subjects with different prognosis into compared groups.	Patients included in the compared groups differ at baseline due to either random events or preferred selection of one type of subjects to the experimental group	Overestimation of the effect of the placebo in studies only involving placebo vs. no-treatment. Unclear impact on the estimated effect of placebo in studies involving active vs placebo vs no-treatment.
Response bias	The tendency for patients or experimental subjects to report their symptoms in a way they feel is socially acceptable or desirable.	Patients or experimental subjects in the placebo group may report symptoms more optimistically than in the no-treatment group	Overestimate placebo effects of patient reported outcomes, for example pain and nausea
Co-intervention bias	The tendency for patients or experimental subjects to seek out and get treatment that is not part of the trial or the experiment.	Patients or experimental subjects in the no-treatment group may be more inclined to seek out non-protocolised interventions	Underestimate placebo effects when the non-protocolised intervention has a clinical effect, either due to a placebo effect or a non-placebo effect
Attrition bias	The tendency for patients or experimental subjects to drop out of the trial or the experiment.	Patients or experimental subjects in the no-treatment group may be more inclined to drop out	Unclear. The degree of bias and its direction depend on whether those leaving the no-treatment group had better or worse outcomes than those who stayed.
Outcome reporting bias	The tendency in scientific publications for statistically significant outcomes to be selected for reporting more frequently than outcomes with insignificant results	The authors of scientific publications often report only a subset of the outcomes studied, and tend to select those with statistically significant results	Overestimate placebo effects in articles aimed at studying placebo. Unclear impact on articles aimed at studying an active intervention (typically active vs placebo vs no-treatment)
Publication bias	The tendency for scientific publications with a statistically significant result to be published more frequently than studies with an insignificant result	Published scientific studies often reflect only a subset of the studies conducted, and those published tend to report statistically significant results	Overestimate placebo effects in articles aimed at studying placebo. Unclear impact on articles aimed at studying an active intervention (typically active vs placebo vs no-treatment)

Causal indeterminateness bias	A placebo intervention will often serve as a 'surrogate' causal factor for the largely indetermined true causal factors	The causal factors of the placebo effect are not typically imbedded in the placebo intervention per se, but in the patient-provider interaction	Competing interpretations of which causal factors are most important in a study finding large effects of placebo would typically have very different clinical implications
Nonclinical settings in laboratory experiments	A laboratory experiment will differ from the typical clinical situation in important ways	Non-clinical experimental studies on placebo tend to be of very short duration and may involve healthy volunteers	Provide valuable insight into the neurobiology and mechanisms of placebo effect, but results cannot reliably be extrapolated to a clinical setting
Informed consent and randomization	The trial or experiment may interact with the patients included	Informing patients about being part of a trial or experiment may alter preconceptions and beliefs	May underestimate or overestimate placebo effects. Beliefs in the effect of an interventions may be less pronounced compared with a clinical situation

Adapted from "Placebo effect studies are susceptible to response bias and to other types of biases. Main types of challenges to the reliability and generalizability of randomized trials and experiments assessing the effect of placebo," by A. Hróbjartsson, T. J. Kaptchuk, and F. G. Miller, 2011, *Journal of Clinical Epidemiology*, *64*(11), p. 1126. Copyright 2011 by Elsevier.

Appendix D: The Preparation Phase

The preparation phase will comprise of two steps.

Step 1 – A pilot study. The pilot study is important in order to establish the equivalence of the chosen placebo cream descriptions. First, two descriptions of the intended placebo creams will be developed. These descriptions will contain comparable, but differently stated information (such as general information about the producer, or the appearance of the cream). The difference would be most pronounced in the intended mechanism of action. For example, in one case, the cream might be described as *"creating a protective cooling layer"*, while the other one could be described as *"cooling by desensitising thermoreceptors in the skin"*. One leaflet will be created for each product and both leaflets will be presented together. For each participant, the location (right or left side) of the leaflets relative to each other would be randomised. Participants would be approached by volunteers in public within cities where the study would take place.

Participants will be first asked to imagine they are about to undergo SPT where they develop a small wheal, then presented with the leaflets, and subsequently asked to rate the two creams based on three questions:

1. How attractive is this cream for you?
2. How effective do you expect the cream to be?
3. If you could choose only one of the creams, which one would you prefer?

The first two questions would be rated on a scale of 1 – 10 with 10 being the most attractive/effective. Participants will be told in advance that they might be asked to imagine an uncomfortable scenario involving skin irritation.

Step 2 – Physician Training. In order to better account for the effect of perceived warmth and competence, physicians administering SPT will undergo a short training programme for the purpose of the study. The aim is that the physician is perceived as competent and moderately warm, as high or low warmness might be a confounder for the purpose of BIS/BAS testing. Behavioural and environmental cues described in the original study will be used (Howe et al., 2017). The specific content of the training would have to be constructed based on the number of medical facilities involved and their respective environment.

Appendix E: Measurement Protocol

Obtaining the measures will be performed as follows:

- **wheal size:** a transparent ruler for allergy testing will be used to measure the mean diameter of the wheal in millimetres. In an instance where the wheal takes on an irregular shape, a mean diameter will be computed as an average of the longest and the shortest perpendicular axes from the centre.
- **flare:** an equivalent method will be used for the flare size. In both of these cases, the methods were selected based on the recommendations from the manual of the Australian Society of Clinical Immunology and Allergy (ASCIA, 2020).
- **itchiness:** participants will be asked to mark their itchiness on a continuous VAS scale where the right end of the line represents the worst itch and the left end of the line no itch at all. The VAS measures will be taken anonymously using an electronic device (such as a tablet) in order to reduce responding bias.

When administering the VAS scale, the physician would inform the participant that they will not see the patient's results for the VAS measures and that all of the measures will be evaluated by an independent researcher in order to reduce bias. In line with that, they will emphasise the importance of honest reports so that the study results are conclusive. The physician would step away and allow the participant to fill in the VAS measure. The wheal size ratings and flare would be submitted by the doctor within the same electronic device without the patient seeing the results.

Appendix F: Data Analysis

The two questions from the preparation phase (the attractiveness of the two creams and their expected effectiveness) will be analysed using dependent samples t-tests (one for each question) or their nonparametric or robust counterparts in the event of an assumption violation. The third question (the preference of one cream or the other) will be analysed using a binomial test. The goal is to have equally compelling options to choose from. Of course, absent evidence against the null hypothesis should not be interpreted as evidence of absence. Rather, the statistical tests will be used as a rough guiding principle, but data exploration will be taken into account.

As has already been mentioned in the main section, the main method of analysis will be two multiple regressions[4] using the following variables: placebo (0/1), choice (0/1), BAS, BIS, placebo x choice, placebo x choice x BAS, placebo x choice x BIS. Because the research hypotheses are directional, the alternative statistical hypotheses will also be one-sided in those respective cases. The continuous predictors will be centred prior to their entry. The following table provides an overview of all chosen predictors and their corresponding research and statistical hypotheses. This table serves as a rough overview of the expected effects but using interaction plots will be necessary for an interpretation. A non-significant lower-order interaction or main effects in the presence of a significant higher-order interaction will not be interpreted as a support against the corresponding research hypothesis without considering the plots. To prevent inflation of type I error, a full model analysis will be reported for testing the main hypotheses even if any of the higher-order terms are insignificant. As part of the exploratory analysis, a minimal adequate model will be identified and tested.

Alpha will be set to 0.05 for each test. Because multiple hypotheses tests will be conducted, the Benjamini-Hochberg procedure will be used to control the false-discovery rate. In order to test the research hypotheses, formal tests are needed only for some of the terms and therefore, only those will be corrected. The rest will serve as exploratory parts of the analysis, including a third and equivalent regression analysis using flare size as the dependent variable. For the exploratory part, no multiplicity adjustments will be used.

[4] It might be argued that a multivariate regression or a MANOVA might be more suitable given that two dependent variables are being handled. However, because the research hypotheses should be evaluated separately for each of the measures given the controversies around the subjective and objective measures, conducting separate multiple regressions with multiplicity adjustment is more suitable for this purpose (Grice & Iwasaki, 2009).

predictor	research hypothesis / exploratory	null	alternative
intercept	exploratory	$\beta_1 = 0$	$\beta_1 \neq 0$
placebo	H1: The administration of a placebo cream decreases the wheal size / VAS reports.	$\beta_2 \leq 0$	$\beta_2 > 0$
choice	exploratory	$\beta_3 = 0$	$\beta_3 \neq 0$
BAS	exploratory	$\beta_4 = 0$	$\beta_4 \neq 0$
BIS	exploratory	$\beta_5 = 0$	$\beta_5 \neq 0$
placebo x choice	H2: The combined effect of choice and administration of a placebo cream decreases the wheal size / VAS reports more than the sum of the individual effects.	$\beta_6 \leq 0$	$\beta_6 > 0$
placebo x choice x BAS	H3: The combined effect of choice and administration of a placebo cream decreases the wheal size / VAS reports more in participants who score higher on the BAS scale.	$\beta_7 \leq 0$	$\beta_7 > 0$
placebo x choice x BIS	H4: The combined effect of choice and administration of a placebo cream decreases the wheal size / VAS reports less in participants who score higher on the BIS scale[5].	$\beta_8 \geq 0$	$\beta_8 < 0$

Table F1: Summary of the research hypotheses and their respective null hypotheses

Assumptions of the main analyses will be checked using diagnostic plots from base R (function plot()): the assumption of normality of residuals (Q-Q plot of residuals), homoskedasticity (scale-location plot), linearity (residuals vs. fitted values), and absence of great outliers (outliers with Cook's D > 0.5)[6]. If any of these assumptions are violated, respective robust methods, non-linear methods or transformations would be used depending on the nature and severity of assumption violation. In such a case, deviations from the pre-planned analysis would be reported.

[5] This hypothesis could also be stated as: Higher BIS trait reduces the effect of placebo X choice interaction.

[6] While formal tests of assumptions such as the Shapiro-Wilk test of normality would provide a more straightforward and transparent alternative, relying strictly on these tests can be misleading with respect to their power and special cases of data.

Using G*Power (Faul et al., 2007), the required sample size for achieving an appropriate power was computed. The linear multiple regression option under the t-tests family was selected. The previous study found a medium effect on size of the wheal (Howe et al., 2017). Because the previous study compared the placebo group to a nocebo group, the expected effect might be smaller. While small effects might be practically irrelevant, they would still be valuable for the comparison of the effect on subjective and objective measures. Moreover, effect sizes for the interactions are generally smaller. As a compromise between a small effect size and an unrealistic sample size that would be required, f^2 was set to 0.085, that is, between the reference small effect of $f^2 = 0.02$ and a medium effect size of $f^2 = 0.15$ (Cohen, 1988). Power was set to the usual value of 0.8, and the number of predictors to 7. Alpha of 0.00625 was used for the case of the strictest adjustment of $\alpha/8$. The computed sample size is 135 participants. This number was adjusted to 136 as it results in 34 participants per group.