Bachelor's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Social and Educational Sciences
Department of Psychology

Candidate 10147

Bachelor thesis

Validation of non-verbal behavior in vidoes designed for experimental research og placebo and pain

Bachelor's thesis in Psykologi
Supervisor: Hojjat Daniali
January 2022

NTNU
Kunnskap for en bedre verden

Candidate 10147

Bachelor thesis

Validation of non-verbal behavior in vidoes designed for experimental research og placebo and pain

Bachelor's thesis in Psykologi
Supervisor: Hojjat Daniali
January 2022

Norwegian University of Science and Technology
Faculty of Social and Educational Sciences
Department of Psychology

NTNU
Norwegian University of
Science and Technology

Table of Contents

Preface

Firstly, I would like to thank my project supervisor Hojjat Daniali, for allowing us a great deal of freedom, in designing our own “miniature” research-project; through our “secondary aim”, and giving great (and frequent) feedback throughout this project. I doubt any other bachelor-groups have gotten nearly as much feedback, on their written work as we have. During this project I was allowed a lot of responsibility - this is something I am very grateful for – which allowed me to develop new skills. I was responsible for a large part of designing our online survey, which was part of the secondary aim for three out of four groups. I designed and selected a large part of the questions and measures in our primary aim; including the questions regarding *hypothetical pain*, *cream-efficacy*, *fear of pain* and *PANAS*, and created the questionnaire itself, on *nettskjema.no.*

I was also responsible for writing the code; used to randomize participants. I would like to point out that I got help in designing the questionnaire. Nora Trohaug played a crucial role in designing the hypothetical scenario used in our online survey, and Aslak Bakke was responsible for editing the videos used in the survey. Hojjat was also responsible for giving feedback on the questionnaire, and gave frequent suggestions on how we could improve it. Lastly, I was responsible for collecting and transposing the data from both our primary and secondary aim into SPSS-files; available to all of the students. In the SPSS-file in our primary aim: The values for each coder was doublechecked; and missing values were filled in by the respective coders.

Abstract

Our study consisted of two “aims”: 1. *Coding of NB* and 2. *Online survey.* In our primary aim: 15 students were trained to code non-verbal behaviour (NB), through a scheme measuring general impressions. The students coded a set of videos displaying three different “channels” of positive NB (PFE, PTV and PBM¸in addition to a neutral condition (NE); where all channels of NB were kept as neutral as possible. These videos are intended, to be used in experimental research, on the relationship between NB placebo and pain. The results showed that there was an acceptable level of inter-rater reliability between coders.

Some questions were raised regarding the validity of the *PFE-videos;* indicating that they displayed increased PTV, compared to the neutral videos. There were also some questions raised regarding not controlling for possible negative aspects of isolating channels of NB. Despite this, the results suggest that the videos can be validated; and used in future experimental research.

In our secondary aim 80 participants were randomized into four groups, where they were showed one out of four different NBs (PBM, PFE, PTV and NE). Participants rated the mood-states of the actors in the videos using PANAS. Our results showed that all videos were rated as more positive than negative. PFE was rated as more positive compared to NE, and PBM and PFE were rated as less negative. Our results showed that isolated channels of positive NB did not lead to increased negative ratings, and supported the findings from our primary aim.

Introduction

Non-verbal behaviour (NB) forms an essential part of human interaction, and communication. NB can be defined as all actions which are not conveyed using words. (Ambady & Weisbuch, 2010). NBs can be divided into micro- and macro-level NBs. Micro-NB's traditionally refers to isolated behaviours, such as smiling, raising eyebrow and forward lean. Macro-NBs generally refers to constellations of micro-behaviours, which together are associated with general psychological meanings, such as warmth and aggression (Ambady & Weisbuch, 2010)

The same sentence can have a widely different meanings, depending on what NB's accompanies it. NB can change the sentence "*I love you"* from a sincere expression, to becoming a sarcastic insult. NB's do not only play a role in our personal relationships, but also when communication with our healthcare-providers. Regardless of what our primary care physician says, their NB's can override the meaning of their words (e.g., Philipot et al., 2003) and this might influence the patients' expectations regarding the treatment, and its outcome (Constantino et al., 2011). Chen et al., 2019 has suggested that NB may be an important medium, for transferring positive expectations from the healthcare provider to the patient; shown through increased placebo-effects. The placebo- and nocebo-effect refers to positive- versus negative effects induced by medically inactive element (McQuay, 2005, Colla et al., 2008).

In this study we test the reliability and validity of a set of videos, intended to research the effects NBs, on placebo and pain. In these videos NBs are isolated into channels, to research the importance of different channels of NB, on the placebo-effect on pain. Daniali & Flaten 2019 reported in a systematic review, that in general; positive NB is associated with lower pain-reports and higher pain-thresholds. The opposite was reported for negative NB; being associated with lower pain thresholds, and higher pain-reports. These effects may also play a role in more general healthcare settings (Colla et al., 2008). Positive NB in experimenters/healthcare providers is for example also associated with less narcotic use, and better emotional- and physical states (Daniali & Flaten, 2019). The placebo/nocebo-effects are also of general relevance in healthcare practices, not only regarding pain. It is also thought to play important roles in both medical, and psychological treatments (Mitsikosta et al., 2011); even if the treatments have "real effects" (Chen et al., 2019).

So far there has been a limited number of studies, looking into the relationship between clinicians/experimenters NBs and placebo/nocebo-effects. Daniali & Flaten (2019) only identified seven studies looking into this; and only two studies researching the effect of specific (micro-) NBs. A Better understanding of the relationship between NB and the placebo/nocebo-effect, could lead to better training of healthcare professionals, and a better understanding of NB as a confounding factor in pain, and treatment-experiments (Daniali & Flaten, 2019). Validating our set of videos, is a first step in bigger research project, looking at the effect between specific NB and the placebo/nocebo-effect.

Three sets of (positive) micro-level NBs were designed (PFE, PBM and PTV) for the videos. These are supposed to represent different "channels" of NB, where the intended NB-channel has been enhanced while the other NB channels have been kept as neutral as possible; only being slightly positive. A fourth macro-NB was designed, where all channels were kept as neutral as possible; only being slightly positive.

If the PFE-video condition also displays a "substantial portion" of PBM - beyond the "slightly positive" in the neutral vide-sections - we cannot assess their individual effects. If we make sure that the videos display the intended NB, they can tell us about the importance of different channels of NB. To assess this, 15 psychology-students were trained; and coded the NB of the videos. Afterwards coded NB was compared between videos and actors. For the primary aim the hypotheses are:

1. Neutral video-sections (Introduction, calibration, pre/post-test and NE) should not be different from each other, on any measures of NB.
2. *PBM* should have higher scores of *gesturing* than all other video-sections, and higher scores of *general impression of positivity* and *expressive*, than all neutral video-sections. PBM should not differ in *positive/friendly tone of voice*, *eye-contact* and *smiling¸* from any of the neutral videos.
3. PFE should have higher scores of *smiling* and *eye-contact* than all other video-sections, and higher scores of *general impression of positivity* and *expressive,* than all neutral video-sections. PFE should not differ in *positive/friendly tone of voice* and *gesturing¸* from any of the neutral videos.
4. PTV should have higher scores of *friendly/positive tone of voice* than all other video-sections, and higher scores of scores of *general impression of positivity* and *expressive,* than all neutral video-sections. PTV should not differ in *smiling, eye-contact* and *gesturing¸* from any of the neutral videos
5. Actors should not differ in coded NB, only in coded *attractiveness.*

When isolating channels of NB, we reduce the ecological validity of the NBs. It may be that PBM is not positive in of itself, but is only perceived as positive when it is displayed in congruence with other channels of positive NB (Stiff et al., 1990). It may be that isolated channels of positive NB, create unintended macro-level effects. When we isolate channels of NB we create a potential incongruency between channels of NB, where they do not correspond with each other. Stiff et al., (1990) for example, suggests that the "mixed messages" of incongruency could be perceived as "deceitful". To assess this, we conducted an online survey, where participants rated the mood-states of the actors in these videos*;* using the PANAS questionnaire (Thompson, 2007). For our secondary aim our hypothesis were:

1. PFE, PTV and PBM is associated with increased ratings of Positive Affect (PA), compared to NE
2. PFE, PTV and PBM is associated with reduced ratings of Negative Affect (NA) compared to NE
3. Mean-scores of PA will be higher than NA, for all videos, including NE.

Primary Aim Coding of NB

Methods; Coding NB

Development of NB; Coding NB

The four sets of micro-NB were developed based upon consistent findings, within the field of NBs. 1) In the PTV-condition actors speak in a warm, strong, friendly, expressive, energetic, and loud tone of voice. 2) In the PFE-condition actors nod and smile frequently, displaying > 5 minutes looking straight at the viewer, more “positive” eyebrow-movements, and display “affirmative” blinking. 3) In the PBM-condition actors lean forwards frequently, reducing the distance between the viewer (camera), to about half a meter. The actors also use expressive and elaborative hand gestures. This includes *indexing, numerical counting with their hands/fingers, affirming*, and *showing/simulating sizes and timelines.* 4) The NE-condition was used as a control-condition. The NE-actors displayed (relatively) neutral behaviours in all the mentioned behaviours. For example, they showed plain faces, monotonous tone of voice, and no gesturing with their hands.

Video-sections; coding NB. The videos were designed to be used in a pain-experiment, consisting of five phases; 1. *introduction, 2. calibration,3. pre-test, 4. conditioning*, and 5. *post-test.* When put together; the videos last about 40 minutes. In the introduction general information and guidelines about the experiment is provided to participants. In the calibration the average pain intensity of participants is calculated individually by giving participants three ascending heat stimulations. In the pre-test participants are given the individually calibrated heat painful stimulus, with no placebo treatment. The experimenter in the video guides the participant through the application, of the painful stimulus.

In sections 1-3 the NB is always kept neutral. In the conditioning section the participants receive a weaker painful-stimulus, associated with a placebo cream; introduced as a powerful heat pain relieving cream. Participants are made to believe that they will receive the same painful stimulus, and are not informed that the cream has no “real” effect. This allows the placebo-cream to be conditioned with less pain. During the conditioning section the placebo-cream and the lowered pain-stimulus is presented with one, out of the four NBs. This is the only section where NB is varied, between participants. The goal of the experiment is to measure the effect different channels of NB has on the conditioning-process. In the last section *the post-test*; participants perform the same procedure, with the same video, pain-stimulus and placebo-cream, as in the pre-test. If one of NB-channels influences the conditioning-process, it should be shown in reported pain, during the post-test.

Three actors performed videos for all five sections. The participants will be shown two of these actors, when shown the whole 40-minute video. This is done to control for experimenters' variability in NB. Actor one informs about the general procedures of the experiment, in the *introduction* and *calibration.* Actor two will inform a about the experimental condition and the placebo treatment, in the *pre-/post-test* and *conditioning.* The actors recorded one version each, for the sections with only neutral NB, and four versions for the conditioning section. These four versions differed only in NB. This together makes 21 videos in total.

Actors; Coding NB

Three female actors were recruited to act in the videos. The actors were cast to fit common stereotypes for healthcare personnel: Caucasian females, slim, in their late twenties, slightly above average height, wearing a small amount of makeup, and dressed in white lab coats (Bridges 1990) NB's have been conveyed successfully through videotapes in prior studies (e.g., Ruben et al., 2017). The actors were trained to display four sets of NBs. *These were positive tone of voice* (PTV), *positive facial expressions* (PFE), *positive body-movement* (PBM) *and neutral expression* (NE). The actors received 10 hours of training, by an expert in NB.

Coders; Coding NB

15 coders were assigned to code the videos. The coders were made up of bachelor-students writing their thesis, on the validation of these videos. The participants varied from 21-25 years old. There were 4 men, and 11 women. Average age was 28.80 (*SD = 1.28)*

Measures; coding of Experimental Videos

The coders used a predefined coding scheme, to code the NB in the video-segments. The scheme measured a predefined set of NBs and general impressions. These were rated using a 9-point scale, for each of NBs and impressions. The scale ranged from "*not at all"* (1) to *"extremely"* (2). The included measures were *smile, eye-contact, gestures, expressiveness, dominant and in charge, friendly/positive tone of voice, overall impression of positivity,* and *attractiveness. Attractiveness* was the only measure, not regarding NB. Ratings of impressions have been shown to be valid and reliable, in trained coders (Blanch-Hartigan et al., 2018) All of these were rated for all 21 video-segments, except for attractiveness which was coded once for each actor. Procedure; coding of NB.

Participants were recruited based on convenience. They were made up by psychology students writing their bachelor thesis, on the validation of these videos. The participants were trained in coding, over a period of 7 weeks. The coders were first educated on NBs, and spent 6 weeks receiving weekly lectures, and researching coding of NB on their own. At week 7 they received training on how to code; using a predefined coding-scheme. This training was given by an expert in NB. The project did not need ethical approval, since it did not deal with any health-related data and it was anonymous; not needing to apply for ethical approval (Helseforskningsloven, 2009, § 2-4). However, the project still conformed with Helsinki ethical codes of conducting research on human beings (World Medical Association, 2013).

The coders coded 21 videos in a fully crossed design; where all coders coded all videos. These videos were thin-slices, roughly 3 minutes long. These thin-slices were cut together segments, of longer videos taken from the beginning, middle and from the end of the full video episodes. The use of thin-slices allows for a more efficient coding process, where less total footage has to be coded (Blanch-Hartigan, 2018). This type of thin-slicing has been shown to be a valid method, for coding NB (Murphy, 2005)

Statistical Analyses; Coding NB

SPSS version 28 (*IBM SPSS Statistics 28*) was used to analyse the data. To measure internal reliability the interclass correlation coefficient (ICC) was used. ICC is one of the most used statistical analysis, for inter-rater reliability (IRR) for scale-, ordinal- and ratio-level measures (Hallgren, 2012) A mixed two-way model was used. We chose to look at consistency among coders, rather than absolute agreement; which is generally advised when computing IRR regarding coding of NB (Blanch-Hartigan 2018). We report the average measure ICC. This is an appropriate fit, seeing that we use a fully crossed design, and our analysis is based on mean-values (Koo & Li, 2016).

To compare similarities and differences in NB between videos, we used a one-way ANOVA. Coding-values were summed across coders. A Tukey post-hoc was used to compare the difference between videos. Since our factor-variable has many levels, it is important to use a post-hoc test that controls for the familywise error (Matsunaga, 2007). An LSD-post hoc would for example not control for the familywise-error rate (Hayter, 1986). (See Appendix C for calculation of familywise error-rate). Tukey-HSD and Bonferroni are both considered robust and conservative tests, which are good at controlling for familywise error rate (Field, 2018). Bonferroni however tend to lose “too much” statistical power when there is a large number of comparisons (Field, 2018). Seeing that we do 147 comparisons across video-sections; Tukey was considered as the best option.

Attractiveness was compared between actors, using a one-way ANOVA. An LSD-post hoc was used to compare the differences between actors. This was chosen due to the low number of comparisons; Stevens (1999) argues that LSD-tests give valid results if there are no more than 3 groups, and there is a significant ANOVA. It was also chosen because we should be more concerned with beta-errors (false negatives) in this case, rather than alpha-errors. This is because attractiveness could be a confounding factor, when using the videos in an experimental setting (Patterson, 1985).

Data screening; Coding NB

Missing values in the data were filled in, and checked by the coders in question. The coders were instructed not to look at other coders ratings. Levene's test for equality of variances, was used to check the differences in variance, between different video-sections. We decided to look at differences based on median, since the data was not normally distributed; Levene's test performs better on non-parametric data if it is based on median, compared to mean. If Levene's test was significant, we could still use an ANOVA; but we would have to use a non-parametric post-hoc (Field, 2018). Levene's test was however not significant.

Results; coding NB

The ICC results (Table 1.1) showed that all coded measures of NB had ICC values, indicating "excellent agreement" (ICC ≥ .90; Koo & Li, 2016), except for dominance, which had an ICC-value of .73, indicating "moderate agreement" (ICC between .50-.75; Koo & Li, 2016). A one-way ANOVA (Table 1.2) showed that there was a significant difference in coded NB, between at least two video-sections; for all coded measures of NB. A Tukey HSD-post hoc showed that there were 57 significant differences total. Means and Standard-deviations are reported in both Table 1.3 and Table 1.4. In addition to reporting mean-differences in text, a table is also included (Table 1.4), to make the results easier to read (American Psychological Association, 2020).

In *gesture:* PBM had a higher score than all other video-sections; differences ranging from ΔM = 80.33-77.67, all differences were significant at p < .001. There were no other significant differences.

In *smile:* PFE had a higher score than all other video-sections; differences ranging from ΔM = 59.33-72.67, all differences were significant at p <.001. There were no other significant differences.

In *eye-contact*: PFE had a higher score than all other video-section; differences ranging from ΔM = 53.00-81.33(PBM-pre-/post-test), all differences were significant at p <.001. There were no other significant differences.

In *friendly/positive tone of voice*: PTV had a higher score than all other video-sections; differences ranging from ΔM = 26.33-54.67(PFE-NE) all differences were significant at p <.001. PFE scored significantly higher than all video-sections; except for PTV. The differences ranged from ΔM = 17.33-28.33 (pre-/post-test-NE), all differences were significant at $p < .05$; expect for the difference between PFE and NE, which was significant at $p < .001$. There were no other significant differences.

In *dominant and in charge:* PFE scored higher than introduction, pre-/post-test and NE. The differences ranged from ΔM = 13.33-15.00 (NE-pre-/post-test). The difference between PFE and NE was significant at $p = .028$, the rest were significant at $p < .01$. PBM scored higher than introduction, pre-/post-test and NE. The differences ranged from ΔM = 12.67-14.33 (NE-pre-/post-test). All differences were significant at $p < .05$. There were no other significant differences.

In *overall impression of positivity*: PFE scored higher than all other video-sections; except for PTV. The differences ranged from ΔM = 23.00-47.67 (PBM-NE). All differences were significant at p < .001, except for the difference with PBM; which was significant at p = .044. PTV scored higher than introduction, calibration, pre-/post-test and NE. The differences ranged from ΔM = 25.00-35.33 (Introduction-NE). The difference with the introduction was significant at $p < .05$. The differences between PTV with calibration and -pre-/post-test were significant at $p < .01$, while the with NE was significant at $p < .001$. PBM scored higher than NE ΔM = 24.67, $p = .028$. There were no other significant differences.

In *expressive:* PBM scored higher than all video-sections; except for PFE. The differences ranged from ΔM = 17.67-45.00 (PTV-NE). All differences were significant at p < .001, except for the difference between PBM and PTV, which was significant at $p = .031$. PFE scored higher than the introduction, calibration, pre-/post-test and NE. The differences ranged from ΔM = 26.67-34.33 (Introduction-NE). All differences were significant at p < .001. PTV scored higher than the introduction, calibration, pre-/post-test and NE. The differences ranged from ΔM = 19.67-27.33 (Introduction-NE). The difference between PTV and the Introduction was significant at p = .015. The difference between PTV and the Pre-/post-test was significant at, p = .007. The rest were significant at p < .001. There were no other significant differences.

A one-way ANOVA showed no significant differences in coded NB, between actors (Table 1.3). However, there was a significant difference in attractiveness, between at least two of the actors, F(2, 19) = 3746.23, p < .001. An LSD-post hoc showed that there was a significant difference in attractiveness, between all the actors. The biggest difference was between actor one (M = 86.75, SD = 0.71) and -three (M = 52.43, SD = 1.13), ΔM = -34.32, p < .001, followed by the difference between actor two (M = 73.00, SD = 0.00) and -three, ΔM = -34.32, p < .001, the smallest difference was between actor one and -two, ΔM = -20.57, p < .001. Actor one was coded as the most attractive, followed by actor two, and la

Table 1.1

*Interclass Correlation Coefficients for Coded N, Using Average Measures and Consistency among coders (*N *= 15)*

Variable	**ICC**^{a}	**95% CI**^{b}
1. Gesture	**.99*****	.98-.99
2. Smile	**.99*****	.98-1
3. Eye contact	**.99*****	.98-.99
4. Friendly/positive tone of voice	**.96*****	.93-.98
5. Dominant and in charge	.72***	.50-.87
6. Overall impression of positivity	**.96*****	.93-.98
7. Expressive	**.96*****	.93-.98
8. Attractiveness	**.97*****	.95-.99

Note: Bold highlights ICC values greater than .90 ("excellent agreement"; Koo & Li, 2016), ***p < .001. a = Inter class correlation coefficient, b = 95% confidence interval

Table 1.2

One-Way ANOVA Coded NB for Video-Sections

Measure	Introduction		Calibration		Pre/post-test		Ne		PBM		PFE		PTV		F(6, 14)	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*		
Gesture	21.67	2.52	20.33	1.15	16.67	0.58	19.67	1.15	**99.33**	9.07	19.33	1.15	17.00	1.00	206.09***	.99
Smile	25.33	7.02	24.00	6.56	22.33	5.86	20.00	7.00	29.00	13.23	**92.67**	16.86	33.33	13.58	16.93***	.88
Eye contact	55.67	12.66	45.00	17.44	42.33	9.29	47.00	7.21	70.67	6.66	**123.67**	3.51	58.67	11.02	21.60***	.90
Friendly/positive tone of voice	46.67	6.81	46.00	6.08	48.33	8.39	37.33	5.13	50.00	6.08	**65.67**	7.23	**92.00**	0.00	26.18***	.92
Dominant and in charge	46.00	3.61	42.33	1.15	42.00	3.46	43.67	3.21	**56.33**	9.07	**57.00**	3.61	45.33	1.15	6.48**	.74
Overall impression of positivity	42.33	10.02	40.33	8.08	37.33	7.51	32.00	5.20	**56.67**	9.61	**79.67**	10.26	**67.33**	3.21	14.22**	.86
Expressive	34.00	7.55	27.00	1.73	32.00	3.46	26.33	3.51	**71.33**	11.15	**60.67**	3.51	**53.67**	4.62	28.33***	.92

Note: ***p < .001, **p < .01, mean-values significantly higher than others; are highlighted in Bold (Tukey HSD), n = 3; for all video-sections.

Table 1.3

One-Way ANOVA Coded Measures for Actors

Measure	Actor one		Actor two		Actor three		F(6, 14)	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*		
Gesture	32.43	33.86	29.86	30.07	29.43	27.24	0.02	.00
Smile	31.71	22.22	31.71	26.28	42.29	31.36	0.36	.04
Eye-contact	65.57	31.21	54.71	30.22	69.57	25.38	0.49	.05
Friendly/positive tone of voice	54.00	18.79	51.14	19.98	60.29	17.14	0.44	.05
Dominant and in charge	50.00	9.63	45.29	5.06	47.29	6.05	0.76	.08
Overall positivity	50.57	16.99	44.86	18.94	57.00	18.64	0.78	.08
Expressive	45.29	21.95	41.71	16.95	43.71	16.89	0.06	.01
Attractiveness	86.71	0.76	73.00	0.00	52.43	1.13	3367.38	.99

Note: ****p < .001, n = 7; for all video-sections.*

Table 1.4

*Tukey post-hoc for differences in coded NB, between video-sections (*N *= 21)*

				Mean difference I – J						
NB	Variables^{I}	M	SD	1^{J}	2^{J}	3^{J}	4^{J}	5^{J}	6^{J}	7^{J}
Gesture	1. Calibration	20.33	1.15	-						
	2. Introduction	21.67	2.52	1.33	-					
	3. Pre-/post-test	16.67	.58	-3.67	-5.00	-				
	4. NE[a]	19.67	1.15	-.67	-2.00	3.00	-			
	5. PBM[a]	99.33	9.07	79.00***	77.67***	82.67***	79.67***	-		
	6. PFE[a]	19.33	1.15	-1.00	-2.33	2.67	-.33	-80.00***	-	
	7. PTV[a]	17.00	1.00	-3.33	-4.67	.33	-2.67	-82.33***	-2.33	-

NB	Variables^{I}	M	SD	1^{J}	2^{J}	3^{J}	4^{J}	5^{J}	6^{J}	7^{J}
Smile	1. Calibration	24.00	6.56	-						
	2. Introduction	25.33	7.02	1.33	-					
	3. Pre-/post-test	22.33	5.86	-1.67	-3.00	-				
	4. NE[a]	20.00	7.00	-4.00	-5.33	-2.33	-			
	5. PBM[a]	29.00	13.23	5.00	3.67	9.00	-63	-		
	6. PFE[a]	92.67	16.86	68.67***	67.33***	70.33***	72.67***	63.67***	-	
	7. PTV[a]	33.33	13.58	9.33	8.00	11.00	13.33	4.33	-59.33***	-

NB	Variables[I]	M	SD	1[J]	2[J]	3[J]	4[J]	5[J]	6[J]	7[J]
Eye contact	1. Calibration	45.00	17.44	-						
	2. Introduction	55.67	12.66	10.67	-					
	3. Pre-/post-test	42.33	9.29	-2.67	-13.33	-				
	4. NE[a]	47.00	7.21	2.00	-8.67	4.67	-			
	5. PBM[a]	70.67	6.66	25.67	15.00	28.33	23.67	-		
	6. PFE[a]	123.67	3.51	78.67***	68.00***	81.33***	76.67***	53.00***	-	
	7. PTV[a]	58.67	11.02	13.67	3.00	16.33	11.67	-12.00	-65.00***	-

NB	Variables[I]	M	SD	1[J]	2[J]	3[J]	4[J]	5[J]	6[J]	7[J]
Friendly/positive tone of voice	1. Calibration	46.00	6.08	-						
	2. Introduction	46.67	6.81	.67	-					
	3. Pre-/post-test	48.33	8.39	2.33	1.67	-				
	4. NE[a]	37.33	5.13	-8.67	-9.33	-11.00	-			
	5. PBM[a]	50.00	6.08	4.00	3.33	1.67	12.67	-		
	6. PFE[a]	65.67	7.23	19.67*	19.00*	17.33*	28.33***	15.67	-	
	7. PTV[a]	92.00	.00	46.00***	45.33***	43.67***	54.67***	42.00***	26.33***	-

NB	Variables[I]	M	SD	1[J]	2[J]	3[J]	4[J]	5[J]	6[J]	7[J]
Dominant and in charge	1. Calibration	42.33	1.15	-						
	2. Introduction	46.00	3.61	3.67	-					
	3. Pre-/post-test	42.00	3.46	-.33	-4.00	-				
	4. NE[a]	43.67	3.21	1.33	-2.33	1.67	-			
	5. PBM[a]	56.33	9.07	14.00*	10.33	14.33*	12.67*	-		
	6. PFE[a]	57.00	3.61	14.67**	11.00	15.00**	13.33*	.67	-	
	7. PTV[a]	45.33	1.15	3.00	-.67	3.33	1.67	-11.00	-11.67	-

NB	Variables[I]	M	SD	1[J]	2[J]	3[J]	4[J]	5[J]	6[J]	7[J]
Overall impression of positivity	1. Calibration	40.33	8.08	-						
	2. Introduction	42.33	10.02	2.00	-					
	3. Pre-/post-test	37.33	7.51	-3.00	-5.00	-				
	4. NE[a]	32.00	5.20	-8.33	-10.33	-5.33	-			
	5. PBM[a]	56.67	9.61	16.33	14.33	19.33	24.67*	-		
	6. PFE[a]	79.67	10.26	39.33***	37.33***	42.33***	47.67***	23.00*	-	
	7. PTV[a]	67.33	3.21	27.00**	25.00*	30.00**	35.33***	10.67	-12.33	-

NB	Variables^{I}	M	SD	1^{J}	2^{J}	3^{J}	4^{J}	5^{J}	6^{J}	7^{J}
Expressive	1. Calibration	27.00	1.73	-						
	2. Introduction	34.00	7.55	7.00	-					
	3. Pre-/post-test	32.00	3.46	5.00	-2.00	-				
	4. NE^{a}	26.33	3.51	-.67	-7.67	-5.67	-			
	5. PBM^{a}	71.33	11.15	44.33***	37.33***	39.33***	45.00***	-		
	6. PFE^{a}	60.67	3.51	33.67***	26.67***	28.67***	34.33***	-10.67	-	
	7. PTV^{a}	53.67	4.62	26.67***	19.67*	21.67**	27.33***	-17.67*	-7.00	-

Note: a = conditioning section, it has four different versions, displaying different NB. **p* = .05, ***p* = .01, ****p* < .001, n = 3 for all videos, values were summed across coders, table is based on general APA-guidelines and Nicoli & Pexman, (2010).

Discussion; Coding NB

Our findings showed that we had “excellent agreement” (ICC < .90) between coders, in seven out of eight measures. *Dominant and in charge* was the only measure which did not have excellent ICC-values; showing only “moderate agreement” (ICC between .50 and .75).

none of the neutral video-sections (Introduction, Calibration, pre/post-test and NE) differed from each other in coded NB. PBM had higher scores of *gesturing* than all videos, and higher scores of *expressivity* than all neutral video-sections. *General impression of positivity* was higher than NE, and *dominant and in charge* was higher than all neutral video-sections, except for the introduction. PTV had higher scores of *friendly/positive tone of voice* than all other video-sections, and higher scores of *general impressions of positivity* and *expressivity,* than all neutral video-sections. PTV’s scores of *expressivity* was also higher compared with PMB. PFE had higher scores of *eye-contact* and *smiling* than all other videos. It also had higher scores of *general impression of positivity, expressivity* and *friendly/positive tone of voice* than all neutral video-sections; *general impression of positivity* in addition higher compared with PBM. Lastly PFE scored higher on *dominant and in charge* than all neutral video-sections, except for the introduction. The only difference in measures between actors was *attractiveness.*

Comparing our results with our hypothesis we see that:

1. Neutral video-sections (*Introduction, calibration, pre/post-test* and *NE*) are not different from each other, on any measures of NB; and the hypothesis is fully supported.
2. *PBM* has higher scores of *gesturing* than all other video-sections, and higher scores of *expressive,* than all neutral video-sections. PBM did not differ in *positive/friendly tone of voice, eye-contact* and *smiling¸* from any of the neutral videos. This supports the hypothesis. *General impression of positivity* was only higher than NE; and not the other neutral video-sections. This does not support the hypothesis. In summary the hypothesis is partially supported.
3. PFE has higher scores of *smiling* and *eye-contact* than all other video-sections, and higher scores of *general impression of positivity* and *expressive,* than all neutral video-sections. PFE does not differ in *gesturing¸* from any of the neutral videos. This supports our hypothesis. PFE does however differ in *positive/friendly tone of voice* from all of the neutral videos. This does not support the hypothesis. In summary the hypothesis is partially supported.

4. PTV has higher scores of *friendly/positive tone of voice* than all other video-sections, and higher scores of scores of *general impression of positivity* and *expressive,* than all neutral video-sections. PTV does not differ in *smiling, eye-contact* and *gesturing¸* from any of the neutral videos This supports the hypothesis. In summary it is fully supported.
5. Actors should not differ in coded NB, only in coded *attractiveness;* and the hypothesis is fully supported.

In summary three out of five hypothesis are fully supported, and two are only partially supported. In general, our results indicate that the videos are validated for use in experimental research.

Firstly, our results indicate excellent inter-rater-reliability. This in of itself does not validate the videos; but lends credibility to the data obtained from the coders. There is also good reliability between "intended" NB across video-sections. There is for example no difference in NB, between video-sections displaying neutral NB. This in accordance with there being no difference in NB between actors, indicates that conditions of NB, are (relatively) similar across videos. These (above mentioned) measures indicate "reliability"; but there is also a question of "validity". The videos aren't only reliable, but also "similar and different, in the way they should be*".* The NBs in the videos, was based on established findings, on what characterises positive NB, in specific channels of NB (Daniali & Flaten, 2019, Ambady & Weisbuch, 2010). When our data suggests that our videos differ (the way they should) in accordance with these findings, it suggests that they can be validated.

It is worth mentioning that *PFE* was coded as having more *positive tone of voice* than all of the neutral-video-sections. This could indicate that we were unable to isolate facial expression, from tone of voice. This could cause problems, when trying to untangle the effects of *PTV* from *PFE.* If the *PFE*-condition leads to stronger placebo-effects, when used in experimental research, it may be caused by the combination of *PFE* and *PTV*; and not by channelised *PFE.* There is some research suggesting that smiling modulates the vocal-chords, causing the tone of voice to sound more positive. It may therefore be difficult to separate *PFE* and *PTV* (Campanella & Belin 2007). There is also a possibility that this was caused by one of the actors smiling more in the *PFE* video-section. Even though the ANOVA did not show any differences in NB between actors, it is unlikely that it would get a significant result; if for example actor three smiled more, in only one out of seven videos. It is also worth mentioning that PTV still was coded as having more friendly/positive tone of voice than PFE.

An important measure regarding the validity of our videos is *overall impression of positivity*. As mentioned, a risk of isolating channels of NBs is creating incongruency; which may cause unintended macro-level interpretations. Isolated channels of positive NB may not be seen as positive, but instead as "deceitful" or passive aggressive (Stiff et al., 1990). The measure *overall impression of positivity* may (to a degree) control for these effects. It is for example likely that NB seen as "deceitful" would get lower scores of *general impressions of positivity.* Our results showed that PTV and PFE, were coded as more positive; than the neutral video-sections. This could indicate that these were not perceived as "deceitful" or negative. PBM was only coded as more positive than NE. It is therefore less certain, that incongruency did not cause unwanted effects in PBM.

In a similar fashion to *general impression of positivity* the measure *dominant and in charge* could also control for incongruency, and its effects. Seeing that dominance could be interpreted as negative. Our results suggest that *PFE* and *PBM* were coded as more *dominant and in charge.* On the surface this could be seen as a sign of negativity (Moors & Houwer 2022); created by incongruency. Dominance is however a complicated concept, and could be seen as both positive and negative (Haley & Sidanius 2006). It is therefore difficult to say which aspect of dominance (positivity versus negativity) this indicates. It is worth mentioning that the ICC-value for dominance had lower scores than the other measures; only showing "moderate agreement". This could also have affected the results (Hallgren, 2012).

Limitations; Coding NB

There are numerous limitations regarding validation of these videos. Even though measures of general impressions have been validated (Blanch-Hartigan et al., 2018), it is far from an objective measure of NB, and is suspect to biases and confounding factors (Harrigan et al., 2008). This is further complicated by the coders not being blinded; being fully aware of what (intended) behaviour they were coding, and how it should be rated. It is common practice to blind coders, in psychological research in general (e.g., Meltzoff & Burton 1979); reducing the risk of bias in assessments (Renoult et al., 2016). It has also been done in NB-research (Bonvicini, 2009), although it is not as common. It is possible that not blinding coders could have given us higher IRR/ICC-scores, than if coders were blinded (Renoult et al., 2016). If so this could be caused by bias from the (knowledge) coders.

Another limitation is the use of thin-slices. Even though it has been validated for use in NB-research, it may not be optimal (Murphy, 2005). Using thin-slices could be seen as a continuous trade-off between effectivity and the added informational value, of coding more of the material (Murphy, 2005). Thin-slices are for example unlikely to capture infrequent events, like emotional outbursts (Blanch-Hartigan, 2018). Given that we have not coded every minute of all the videos, we cannot “completely” guarantee the quality of the whole video-segments. This could be seen as a small cost, when compared with the effectivity of using thin-slices.

When validating our videos with channelised NB, one could argue that we at “best” only can validated “compared” to something. In our case this could be our “neutral” videos. For example we “validate” PBM through increased *gesturing,* compared to the neutral videos. The question thereby becomes whether we can validate our neutral videos. The neutral videos are intended to be “neutral” in their NB; only being slightly positive. It is very important that our neutral videos are not “negative”. If neutral in “reality” was negative; the other video-sections being “more positive”; could mean that they in fact only are neutral.

This is complicated since we did not have any measures (in our primary aim), looking at negativity. Even though the neutral video-sections were “slightly positive” on the *general impression of positivity* (*M* = 32.00-42.33), it is not a perfect indication of their positivity/negativity. A prime example of this is PANAS. In PANAS, PA and NA are conceptualised as relatively separate factors (e.g., separate types of behaviours), only having a low correlation (Merz et al., 2013). If this is the case, it could be that we only measured positivity and not negativity, when coding the videos. Even though NE has a *general impression of positivity* mean of 32.00, it could also have a (hypothetical) mean of *general negativity* mean of 40.00. This problem is not only present for the neutral videos, but also for the “positive” videos; where measuring negativity could help control for potential negative interpretations, caused by incongruency.

Strengths; Coding NB

Even though using subjective measures based on impressions can be seen as a limitation, it could also be seen as a strength. While this form of subjective coding could be subject to bias and confounding factors (Harrigan et al., 2008); NBs and its meaning is defined by subjective-, not objective measures (Ambady & Weisbuch, 2010). For example it does not matter if someone's NB is "positive" based on "objective" measures - such as duration and frequency of: *smiling* and *eye-contact* - if the NB is perceived as "negative". In the end, NBs and their meaning, are defined by human interpretations, not by objective measures.

This is of particular interest regarding some of the more complex measures of NB, like *positive/friendly tone of voice* and *general impressions of positivity.* Even though objective measures (in general) often are considered better than subjective ones; one could (in this case) argue the opposite; particularly when seven our of eight measures showed "excellent agreement" among coders. One could make the argument that high ICC-values are less impressive for (relatively) objective measures, compared to subjective ones. It would for example not be very impressive if 5 coders counted the same number of smiles, during a three-minute video-segment. It would however be impressive if all 5 coders "perfectly" agreed on some general impression, like positive tone of voice. Another strength is the use of a fully crossed design, where all coders code all videos (Hallgren, 2012). This is because it is less susceptible to differences between individual coders, making mean-values for the videos more trustworthy. If we for example used a mixed-design – where some of the videos are only coded once, by a single coder – we would have less certain data, for videos coded by a single coder. The "trustworthiness" of our means is also increased by a (relatively) large number of coders.

Secondary Aim; Online Survey

Methods; Online survey

Participants; online survey

100-participants answered the online survey, these were split into four groups; *PTV (n = 25), PBM (n = 25), PFE (n = 25)* and *NE (n = 25).* After removal of participants fulfilling the exclusion criteria, the total number was 80 participants. Participants per group were reduced to; PTV (*n* = 19), PFE *(n* = 22), PBM (*n* = 19), NE (*n* = 20). The average age of all participants was 25.20 (*SD* = 7.70). The age ranged from 15-52 years old. 54 (68%) were female, 26 (33%) were male, and 0 (0%) defined them self as neither male, nor female. The educational level of the participants ranged from *finished 10 years of education*, and *finished or started phd. or equivalent.* The mode education was *finished or started bachelors-degree (Frequency* = 40).

Procedure; online survey

Participants were sampled based on conveniency. Invitations were sent out through various digital platforms. Invitations was sent out in Facebook-groups, and on NTNUs' own digital forum, *innsida.ntnu.no.* These invitations targeted psychology students on NTNU. In addition, students working on the bachelor-project, reached out to fellow students, friends and family. The questionnaire contained 36 questions in total, in addition the participants were also asked to imagine a painful scenario, before watching a short 1-minute video, presenting a fictional pain-relieving cream, named "*embla".* Participants were split into four different groups. Each group was shown a different video. These videos were 1-minute segments from the *conditioning-videos*, cut to fit a hypothetical scenario. The content in the videos were the exact same, only differing in NB: *PTV*, *PBM*, *PFE* and *NE.*

Participants were asked to imagine a painful hypothetical scenario. Where the participants were asked to imagine, that they had burned themselves on a frying pan. The only difference between the questionnaires, was which type of NBs, that were displayed by the actor in these videos. The platform used for the questionnaire was *nettskjema.no.* This platform was developed by UIO (*The university of Oslo)* and made for use in scientific surveys. It has various features ensuring the anonymity, and data-security of the participants (UIO, 2021). The research project did not need to apply for ethical approval, since it did not deal with any health-related data (Helseforskningsloven, 2009, § 2-4).

Randomization; Online Survey

Participants were randomized to one out of four video conditions. The participants were randomised to one out of four groups. This was done through an external link, which randomly redirected the participants to one out of four questionnaires. When groups reached the specified number of participants, the questionnaires were closed, and participants would be redirected to the other groups. The desired group-sizes were 25-, bringing the total to 100 participants. The four questionnaires only differed, in what NB was shown, in the 1-minute video-segment.

Measures; Online survey

The questionnaire contained 36 questions in total. These included; *demographic information, a 10-item personality inventory* (Erhart et al., 2009*), a question regarding digital- work, education and meetings, a shortened version of the fear of pain scale* (Mc Neil et al., 2018), *expected pain-intensity, expected cream efficacy, rated mood-state*, and a *control question,* checking if participants paid attention, when watching the video. In our secondary aim we looked into how participants rated *mood-states,* of the actors in the videotape. *Perceived mood-state* was measured using a shortened 10-item version of the PANAS (Positive and negative affect scale) (Thompson, 2007) PANAS has two dimensions, *Positive Affect* (PA) and *negative affect* (NA). In the short-form version, there is 5 items per dimension. PANAS was developed to measure emotional states in self-report, but has also been validated to measure interpretation of others' mood-states; even being used in research on NB (e.g., Saerbeck and Bartneck 2010).

Data-screening; Online Survey

Exclusion-criteria were; using > 60 minutes to complete the survey, and giving the wrong answer in a control-question. Participants fulfilling the exclusion-criteria, were excluded**.** The datasheet was checked for missing values, none were found. PA and NA were calculated from the 10 items in the PANAS-questionnaire. PA- and NA-scores were checked for normality, within each group, and for the entire sample; the normality was not violated.

Statistical analysis; Online Survey

SPSS version 28 (*IBM SPSS Statistics 28*) was used to analyse the data. Before analysing the data, normality in emotional ratings was checked for each group, and for the sample as a whole. A One-way ANOVAs was performed to analyse the difference in PA and NA affects within and between groups. To compare differences, we used 10 planned contrasts. Bonferroni's correction was used to adjust the significance-level. One for each group comparing NA with PA (4), three comparing PA between: NE-PBM, NE-PTV and NE-

PFE, and three comparing NA between NE-PBM, NE-PTV and NE-PFE. Levene's was significant, *p* = .008. Indicating that we should use planned contrast not assuming equal variances (Field, 2018).

Results; Online Survey

A one-way ANOVA showed that there was a significant difference in NA and PA-between or within, at least two of the groups, *F*(7, 154) = 15.36, *p* < .001. Means and Standard deviations are shown in table 2.1. Planned contrasts revealed that all videos had significantly higher scores of PA compared to NA. The biggest difference in PA and NA was seen in PFE, *ΔM* = 1.64, *p* < .001, followed by: PBM, *ΔM* = 1.26, *p* < .001, PTV, *ΔM* = 1.04, *p* < .001. The smallest difference between PA and NA was seen in NE, *ΔM* = 0.48, *p* = .021. Only PFE had higher scores of PA, compared with NE, *ΔM* = 0.58 *p* = .005. PFE and PBM had significantly lower scores of NA, compared with NE. The biggest difference was between PFE and NE, *ΔM* = -0.61, *p* = .004. The difference in NA between PBM and NE was -.056 *p* = .006.

Table 2.1

Means and Standard-deviations for PA and NA, groups in online survey (N = *80*)

Video	Measure	*M*	*SD*
NE:	PA[a]	2.56	0.67
	NA[b]	2.07	0.56
PFE:	PA[a]	3.15	0.59
	NA[b]	1.51	0.69
PBM:	PA[a]	2.73	0.80
	NA[b]	1.46	0.66
PTV:	PA[a]	2.71	1.09
	NA[b]	1.68	0.74

Note: a = Positive Affect, b = Negative Affect, Measured using PANAS.

Discussion Online survey

Our results showed that all videos had significantly higher scores of PA, than NA. This indicates that all the videos were interpreted as more positive than negative. Only PFE had significantly higher scores of PA compared with NE (Neutral), indicating that it was the only video interpreted as more positive than NE. Both PFE and PBM had lower scores of NA, compared with NE (Neutral). This indicates that they were interpreted as less negative than NE. PTV had neither significantly higher scores of PA-, or lower scores of NA comparted to NE. Looking at our hypothesis we see that:

1. PFE had higher scores of PA compared with NE; PBM and PTV did not have significantly higher scores of PA. This hypothesis is only partially supported.
2. PBM and PFE had lower scores NA compared with NE; PTV did not have significantly lower scores of NA. This hypothesis is only partially supported
3. All videos had higher scores of PA than NA. This hypothesis is fully supported.

Only one out of three hypotheses were fully supported, and two were partially supported. Even though we did not show that all positive videos were interpreted as less negative-, and more positive than NE, we still showed that all videos were significantly more positive than negative. This is an important finding, seeing that we did not control for negativity in our primary aim. It is also worth mentioning that even though PBM and PTV were not rated as more positive, and PTV was not rated as less negative, none of the videos were rated as less positive- or more negative, compared with NE. Although this was not the expected result, it supports the idea that our videos do not display unintended negative macro-NB, created by incongruency. If the isolated NB was perceived as “deceitful”, or “passive aggressive” (Stiff et al., 1990), we would expect them to be rated as more negative, and likely less positive.

Limitations Online survey

If we calculate our statistical power (see Appendix B). We can see that we would need a Cohens d-value of 1.36, indicating that we would need 1.36 SDs difference between the two means - values above 0.8 are generally considered to be strong (Cohen, 1992) - to get significant results. This means that we are likely to miss effects, of moderate or weak sizes (Field 2014). Given that our videos differ in micro-NB, and our measure looks at a macro-level (PANAS), it is likely that we would expect small to moderate effects. It is therefore possible that there are unidentified “real effects”. For example, all the “positive” videos, had higher PA-scores, and lower NA-scores compared with NE, but only three differences were significant. It is possible that some (if not all) of these differences would have been significant, given the right statistical power.

Another limitation is the use of thin-slices; our video-clips being only 1-minute in length. Blanch-Hartigan (2018) suggests that this is on the shorter end, of what is acceptable. It is possible that we would have seen stronger (and "more valid") effect sizes in our analysis; had the videos been longer. These thin-slices may also, not be representative of the full-length conditioning-videos. There are two main reasons for this: Firstly, the videos only showed actor one (out of three). Secondly, they were not composed of randomly selected segments – of the larger videos – but were cut to fit a hypothetical scenario. It is also important to mention that we used a shortened version of PANAS, which – even though validated – is not as good as the full-length questionnaire (Merz et al., 2013).

Beyond this we should also remember that we only looked at the conditioning-videos; and not at the introduction, calibration nor the pre/post-test. We should also be vary of our sample; seeing that it was not randomly selected but based on convenience. Simultaneously it also consisted of a very narrow range of the population; mostly consisting of psychology-students. Seeing that they studied psychology they may also have more knowledge than the general population on: NB, placebo, hypothetical pain and research-designs in general (Henrich et al., 2010). It is possible that this could have affected our results. Because of these limitations we should both be careful in interpreting our results, and in generalising to *all* of the *full-length* videos.

Strengths Online survey

While our online survey has a fair deal of limitations, there are also considerable strengths. First of all, we used a randomized controlled study; where participants were blinded to the NB of the actor. Participants were completely unaware of what the independent variable was; not knowing that we varied NB between groups. They were also told that the main dependant variables were the evaluation of a hypothetical treatment, and hypothetical pain. This minimized potential bias, when assessing the actors through PANAS. It is also worth noting that only using a single actor, removed the potential of *actors* being a confounding variable. Another strength of our design is using PANAS to control for both negativity and positivity. PANAS also measures individual "types of behaviour" such as "hostility" and "nervousness". These could represent potential specific (unintended) macro-level impression, caused by isolating channels of NB. This in addition with the general assessment of both PA and NA; can potentially give good indications of the “unintended” effects of incongruency.

General Discussion

In our primary aim the results (in general) indicate that our videos can be validated, for use in experimental research; even though there were some unwanted/unexpected results. In addition, a question was raised regarding not controlling for negativity. Since some research and measures suggests (e.g., PANAS) that negativity and positivity can be seen as two “relatively” separate factors, we should control for both. This lack of control raised two questions:

1. How can we know if our control-condition is neutral, negative or positive? The validation of our other videos are in large part only validated when compared to this condition. It is therefore important to ensure that the control condition is either neutral or slightly positive, and not negative.
2. When not controlling for negativity we also run the risk of not identifying unintended macro-level effects – such as being interpreted as “deceitful” or passive aggressive – when we isolated channels of positive NB. This could be a result of incongruency, when different channels of NB do not correspond with each other (Stiff et al., 1990.)

As a supplement to our primary aim, we soke to – at least partially – answer this question. In our secondary aim we showed that 1-minute slices of PBM, PTV, PFE and NE: were all interpreted as more positive than negative, when measured using PANAS. This suggests that the isolation of channels of NB, did not lead to negative and unintended interpretations of the NB. Simultaneously we also showed that NE was more positive than negative, which suggests that our control-condition is at least slightly positive. These results on their own do not validate our videos, but is a good supplement to our primary aim; strengthening the validity of the results and the videos themselves.

It is also interesting to note that PFE was the only video to be rated as more positive than NE (in our secondary). This is particularly interesting when comparing results with our primary aim. The results from our primary aim indicated that PFE displayed some *friendly/positive tone of voice.* This could be the reason it is rated as more positive compared with NE, when the other channels of NB were not. PFE was also the only video-section that was rated as more positive than other channels of positive NB (PBM), in our primary aim. This is an example showing that it could be hard to disentangle the effects of the two channels from each other. This may become a problem when applying the videos, in experimental research.

It is also interesting to note that PFE and PBM was rated as less negative than NE, in our secondary aim. Although this does not increase the "absolute" level of positivity, it does increase the level of "relative" positivity; since PA becomes relatively higher compared to NA. If our dependent variable had been the ratio between PA and NA, we may have gotten different results. It may be of interest, when applying these videos in experimental research, to consider what – if we find any effects of NB on placebo and pain – best predicts the effects of these NBs. Is it for example: the reduction in negativity-, the increase in positivity, or the relative relationship between the two? This could be assessed by adding PANAS as a measure, to the experimental procedure.

Implications for further research

This study mainly concerns methodological practices, regarding coding and validation of videos displaying NBs. Our study supports the idea that NB can be coded using: (relatively) simple coding-scheme, trained students as coders and the use of "thin-slices". Further research should however learn from some of our "mistakes". For example, there should be put a bigger emphasis on ensuring the quality of your control-condition, when creating NB-videos. This is not only important when designing/validating videos intended for experimental research, but also for validation in general. In our design one could make the argument (if being strict) that we mainly validated our videos, relative to our neutral conditions.

To control for the "neutrality", it is recommended that one includes measures of both positive, but also negative NBs. As seen in PANAS, these can be seen as two (relatively) independent factors, and it may therefore not be sufficient; to only measure positive aspects of NB. We also advice the consideration of blinding coders. Lastly we have shown that it is possible to supplement coding of NB done by trained coders, with ratings done by "regular" participants. This does not (in any degree) render the coding unnecessary but can act in a supplementary role. It may be particularly helpful, when trained coders are not blinded. It should therefore be considered in further research, when validating videos of NB.

Conclusion

In this study we validated a set of videos displaying different channels of NB, intended to be used in experiments looking at the effect of isolated channels of NB, on pain and placebo. Four sets of NB was designed for the videos. Three of these were supposed to represent isolated channels of positive NB (PBM, PTV and PFE). We also had a “control” where all NB was kept as neutral as possible. In our primary aim, psychology-students were trained in coding NB over a period of 7 weeks. Videos were coded with a coding scheme, using general impressions to measure NB. The students coded “thin-slices” of the videos, in a fully crossed design. Our results indicated that this is an acceptable way of coding videos. Our results largely validated the videos; implying that the videos were successful in isolating channels of NB, while keeping other channels neutral. One exception was PFE, which displayed some signs of PTV. The results also indicated that we were successful in keeping the neutral conditions relatively neutral.

In addition to our primary aim, mood-states of the actors in the videos were rated using PANAS, in an online survey. Our results supported the findings in from our primary aim, indicating that all videos were more positive than negative. In summary the results from both aims, suggest that the videos can be used in future experimental research, looking at the effects of isolated NB on pain and placebo.

References

World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. (2013). *Jama*, *310*(20), 2191-2194. https://doi.org/10.1001/jama.2013.281053

Association, A. P. (2020). *Publication Manual of the American Psychological Association, (2020)*. American Psychological Association.

Bebbington, P. (1995). The content and context of compliance. *International Clinical Psychopharmacology*.

Blanch-Hartigan, D., Ruben, M. A., Hall, J. A., & Schmid Mast, M. (2018). Measuring nonverbal behavior in clinical interactions: A pragmatic guide. *Patient Education and Counseling*, *101*(12), 2209-2218. https://doi.org/10.1016/j.pec.2018.08.013

Bonvicini, K. A., Perlin, M. J., Bylund, C. L., Carroll, G., Rouse, R. A., & Goldstein, M. G. (2009). Impact of communication training on physician expression of empathy in patient encounters. *Patient Education and Counseling*, *75*(1), 3-10. https://doi.org/10.1016/j.pec.2008.09.007

Bridges, J. M. (1990). Literature review on the images of the nurse and nursing in the media. *Journal of advanced nursing*, *15*(7), 850-854.

Brunoni, A. R., Lopes, M., Kaptchuk, T. J., & Fregni, F. (2009). Placebo response of non-pharmacological and pharmacological trials in major depression: a systematic review and meta-analysis. *PloS one*, *4*(3), e4824.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535-543. https://doi.org/10.1016/j.tics.2007.10.001

Chen, P.-H. A., Cheong, J. H., Jolly, E., Elhence, H., Wager, T. D., & Chang, L. J. (2019). Socially transmitted placebo effects. *Nature Human Behaviour*, *3*(12), 1295-1305. https://doi.org/10.1038/s41562-019-0749-5

Chiffi, D., & Zanotti, R. (2016). Knowledge and belief in placebo effect. The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine,

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. Psychological bulletin, Colloca, L., Sigaudo, M., & Benedetti, F. (2008). The role of learning in nocebo and placebo effects. *Pain*, *136*(1-2), 211-218.

Constantino, M. J., Arnkoff, D. B., Glass, C. R., Ametrano, R. M., & Smith, J. Z. (2011). Expectations. *Journal of clinical psychology*, *67*(2), 184-192.

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*(3), 245-265. https://doi.org/10.1348/0144665031752934

Daniali, H., & Flaten, M. A. (2019). A qualitative systematic review of effects of provider characteristics and nonverbal behavior on pain, and placebo and nocebo effects. *Frontiers in psychiatry*, *10*, 242.

Field, A. (2018). *Discovering statistics using IBM SPSS statisitics* (Vol. 5). Sage.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606.

Haley, H., & Sidanius, J. (2006). The Positive and Negative Framing of Affirmative Action: A Group Dominance Perspective. *Personality and Social Psychology Bulletin*, *32*(5), 656-668. https://doi.org/10.1177/0146167205283442

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23.

Harrigan, J., Rosenthal, R., & Scherer, K. (2008). *New handbook of methods in nonverbal behavior research*. Oxford University Press.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the american statistical association*, *81*(396), 1000-1004.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. doi:10.1017/S0140525X0999152X

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pre-/post-test-posttest design: a potentially confusing task. *Psychological bulletin*, *82*(4), 511.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kupers, R., Faymonville, M.-E., & Laureys, S. (2005). The cognitive modulation of pain: hypnosis-and placebo-induced analgesia. *Progress in brain research, 150*, 251-600.

Mast, M. S. (2007). On the importance of nonverbal communication in the physician–patient interaction. *Patient Education and Counseling*, *67*(3), 315-318. https://doi.org/10.1016/j.pec.2007.03.005

Matsunaga, M. (2007). Familywise error in multiple comparisons: Disentangling a knot through a critique of O'Keefe's arguments against alpha adjustment. *Communication Methods and Measures*, *1*(4), 243-265.

McQuay, H. J. (2005). Placebo. *Postgraduate Medical Journal*, *81*(953), 155-160. https://doi.org/10.1136/pgmj.2004.024737

Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature, 282*(5737), 403-404.

Merz, E. L., Malcarne, V. L., Roesch, S. C., Ko, C. M., Emerson, M., Roma, V. G., & Sadler, G. R. (2013). Psychometric properties of Positive and Negative Affect Schedule (PANAS) original and short forms in an African American community sample. *Journal of Affective Disorders*, *151*(3), 942-949. https://doi.org/10.1016/j.jad.2013.08.011

Mitsikostas, D. D., Mantonakis, L. I., & Chalarakis, N. G. (2011). Nocebo is the enemy, not placebo. A meta-analysis of reported side effects after placebo treatment in headaches. *Cephalalgia*, *31*(5), 550-561. https://doi.org/10.1177/0333102410391485

Moors, A., & De Houwer, J. (2005). Automatic processing of dominance and submissiveness. *Experimental Psychology*, *52*(4), 296-302.

Murphy, N. A. (2005). Using Thin Slices for Behavioral Coding. *Journal of Nonverbal Behavior*, *29*(4), 235-246. https://doi.org/10.1007/s10919-005-7722-x

Nicol, A. A., & Pexman, P. M. (2010). *Presenting your findings: A practical guide for creating tables*. American Psychological Association Washington, DC.

Oldridge, N. B., & Streiner, D. L. (1990). The health belief model: predicting compliance and dropout in cardiac rehabilitation. *Medicine & Science in Sports & Exercise*.

Patterson, C. (1985). What is the placebo in psychotherapy? *Psychotherapy: Theory, Research, Practice, Training*, *22*(2), 163.

Philippot, P., Feldman, R. S., & Coats, E. J. (2003). The role of nonverbal behavior in clinical settings: Introduction and overview.

Renoult, J. P., Bovet, J., & Raymond, M. (2016). Beauty is in the efficient coding of the beholder. *Royal Society Open Science*, *3*(3), 160027.

Ruben, B. D., & Gigliotti, R. A. (2017). Communication: Sine qua non of organizational leadership theory and practice. *International Journal of Business Communication*, *54*(1), 12-30.

Saerbeck, M., & Bartneck, C. (2010). Perception of affect elicited by robot motion.

Stiff, J. B., Hale, J. L., Garlick, R., & Rogan, R. G. (1990). Effect of cue incongruence and social normative influences on individual judgments of honesty and deceit. *Southern Journal of Communication*, *55*(2), 206-229.

Testa, M., & Rossettini, G. (2016). Enhance placebo, avoid nocebo: How contextual factors affect physiotherapy outcomes. *Manual therapy*, *24*, 65-74.

Thompson, E. R. (2007). Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology*, *38*(2), 227-242. https://doi.org/10.1177/0022022106297301

Voelkl, B. (2019). Multiple testing: correcting for alpha error inflation with false discovery rate (FDR) or family-wise error rate?

Appendix

Appendix A

The Use of Tables; Reporting Results From ANOVAs and Post-Hoc Tests

In APA 5th it was standard to report ANOVAs and post-hoc tests, in tables (APA, 2001). This was changed in APA 6th (and APA 7th), making it standard to report ANOVAs and post-hoc tests in text, and not tables (APA, 2010). One should however report ANOVAs and post-hoc tests in tables, if there are many ANOVAs, and comparisons (American Psychological Association, 2020; Nicoli & Pexman, 2010); with the goal of making the statistics “understandable to the reader” (American Psychological Association, 2020, p. 181). The table for post-hocs in my primary aim is based on the general APA guidelines for tables presented in (Nicoli & Pexman, 2010). There is no “gold-standard” for post-hoc tables, so I took some creative freedom making the table, and it is my personal design.

In general one should choose only one of the options (reporting either in table or text); in this case my supervisor advised me to include text as well (in addition to my tables).

Appendix B

Calculation of statistical power (sensitivity) of our online survey.

This is a calculation (Table B1) of sensitivity of our planned contrasts, in our online survey. It was calculated using the: participants in PFE and NE, and the Bonferroni corrected significance-level (comparisons = 10) p < .005. The calculation was performed using G*Power (version 3.1.9.7). The results indicate that we would need 1.36 standard-deviations of difference between the two means (Cohens d).

Table B1

Analysis of Statistical Power (Sensitivity), online survey; planned contrasts

Input	α error probability[a]	.005
	Power (1-β error probability)	.95
	Total sample size (N)	80
	Number of groups	4
Output	Noncentrality parameter δ	4.40
	Critical t	2.70
	Df	40
	Effect size d[b] =	1.36

Note: a = same as significance-level, b = Cohen's d

Appendix C

Calculation of familywise error-rate (FWER) and expected alpha-errors.

This is a calculation of expected (FWER) if we used an LSD-post hoc, when comparing differences between coded NB, for video-sections. Given that we have 7 video-sections, we make 21 comparisons, formula:

$$number\ of\ comparisons = n\frac{k(k-1)}{2} = 7\frac{7(7-1)}{2} = 147$$

Note: n = number of tests (ANOVAS/measures)

Given that we have 7 measures of NB in our coding-scheme, this means that we in total have 147. If we do not control for the familywise error, we would expect 7 alpha errors (false positives) (Gigerenzer, 2004), and a >99.9% likelihood of getting at least one alpha error.

$$familiwise\ error\ rate = 1-(1-\alpha)^{n} = 1-(1-.05)^{147} = .99946\ldots$$

Note: n = number of comparisons, α = alpha error rate (.05)

$$\textbf{\textit{expected number of alpha-errors}} = n \cdot \ \alpha = 147 \cdot \ .05 = 7.35$$

Note: n = number of comparisons, α = alpha error rate (.05)

In total we would expect around 7 alpha-errors (false positives). If we combine this with the total number of significant differences in our Tukey; we would find that 7/ 64 positives would be alpha-errors. Our alpha-error rate would therefore be changed from (the standard .05 (5%)) to .11 (11%)

NTNU

Norwegian University of
Science and Technology