Assessing aggression: a physical approach

Jeroen Lennard van Oorschot

Graduation Thesis

Media Technology MSc programme, Leiden University
Thesis advisors: Maarten H. Lamers & Ricardo E. De Gouveia da Costa Cachucho

August 2016

Leiden & Helsinki, Summer 2016

Abstract

While aggression has been studied by many disciplines and in many contexts, the bulk of the research tends to rely on self-report methods. There are multiple problems associated with these methods when applied to contexts where social desirability can introduce a bias. We propose a new, physical research method for aggression research, by measuring how people punch a device under different stimuli. The technical development of the instrument is detailed and the method is tested in a case study on gender aggression. Various considerations and reflections on implementing the method are provided.

1. Introduction

There is no proper restitution for violent behaviour. A stolen car or mobile phone is a nuisance, but it can ultimately be replaced by a similar model. However, a life that is unjustifiably taken cannot be replaced. Not only the victim suffers, but family, friends, and even society does too. Even if the violence has taken place without lethal consequences and the physical wounds have healed, a victim can have sustained psychological trauma for the rest of its life (Leyton, 2003).

It is therefore not surprising that violence and aggression have long been topics of scientific study. Scientists from all disciplines, including psychology, biology, sociology and even mathematics (Johnson, et al., 2011) study aggressive behaviour. Trying to understand the reasons for individuals resorting to violence is predominantly the domain of the social sciences. We propose that the research methodologies to which this domain often constrains itself are inadequate to study the topic of violence satisfactorily. Therefore, we a novel research method that attempts to mitigate some of the problems associated with self-report methods.

First, we make the argument that self-report methods used in social sciences are problematic, when the problem of social desirability is not properly accounted for. Subsequently, we detail the iterative design process of the proposed new research methodology for assessing physical aggression. Finally, we test the proposed methodology with a case study and reflect on its performance.

2. Self-report bias in aggression research

Inherent to self-report methods used in the social sciences, with questionnaires and interviews as prime examples, is the problem of response bias. This is just one of the many measurement artefacts associated with self-report methods, but in the interest of brevity we will focus on the credibility of the answers. For years it has been known that response bias lowers the validity of research (Fiske & Pearson, 1970), yet self-report methods remain dominant in, for example, the field of personality psychology (Robins, Tracy, & Sherman, 2007)

The credibility of the answers given by a respondent bears a direct relation to the validity of the research. The degree to which respondents answer truthfully is affected by a social desirability bias, especially when dealing with socially undesirable topics such as aggression (Saunders, 1991, p. 337). It introduces a bias

when respondents give answers that they think will make a better impression on the researcher, as they are socially more approved.

To assess how aggressive a person is, self-report aggression questionnaires are a commonly used method. These questionnaires focus either on past violent behaviour (for example Archer, Holloway, & McLoughlin, 1995), on how participants think they will behave in hypothetical situations (for example Harris & Knight-Bohnhoff, 1996), or pose questions about how aggressive the participants regard themselves in general (for example Buss & Perry, 1992). Various studies have found correlations between measures for socially desirable answering and aggression measures in these questionnaires, up to $r = -.60$ (Vigil-Colet, Ruiz-Pamies, Anguiano-Carrasco, & Lorenzo-Seva, 2012). The more someone is inclined to answer socially desirable, the more that person tends to underreport aggression (Gregoski, Malone, & South Richardson, 2005; Morren & Meesters, 2002). Social desirability bias also troubles the interpretation of indirect measures of aggression, such as impulsivity, due to the problem of common method variance (Donaldson & Grant-Vallone, 2002; Vigil-Colet, Ruiz-Pamies, Anguiano-Carrasco, & Lorenzo-Seva, 2012). While there are multiple ex-ante and ex-post remedies for common method variance, these are often either not correctly applied or altogether absent (Hult, 2011).

Many authors relying on self-report methods reflect extensively on the biases inherent to their methods and apply (statistical) techniques to counter them. Yet, few make explicit why they adopt these methodologies in the first place. Aggression questionnaires, most notably the aptly dubbed Aggression Questionnaire developed by Buss and Perry (1992), are the de facto standard instruments in the field of aggression research. This is partly because they are very well validated and provide ample data for comparison (see Morren & Meesters (2002) for an example of such a validation and comparison). Furthermore, their quantitative output allows for extensive statistical analyses. Besides, one needs little imagination to see the ethical difficulties associated with studying interpersonal physical aggression in an experimental setting. Statement 22 on The Aggression Questionnaire reads: “If somebody hits me, I hit back.” (Buss & Perry, 1992). It is less troublesome to ask to what degree a respondent finds that statement characteristic of himself, than to test it empirically.

Some research has been done following a more quantitative, observational methodology using proxies such as videogames. For example, Eastin (2006) has

found that the gender of the player and the gender of the opponent modulate aggression. Female players tend to play more aggressively against male characters and, inversely, aggression tends to decrease in male players playing against female characters.

3. Iterative method design

In the previous section, a number of issues associated with self-report methods have been identified, most notably the social desirability bias for undesirable traits. This section details the exploration into a new research methodology that aims to be to a lesser extent affected by this type of bias. It consists of a punchable instrument that is capable of measuring both the number of punches within a certain timeframe as well as their intensity. Such an instrument could be used in a number of contexts. For this study, we chose to develop the method with stimulus genderisation as independent variable, as we will study this concept in the case study.

To the best of our knowledge, such a method has not been applied to this field of research before[1]. Nor were we able to find an off-the-shelf device capable of reliably measuring punches in quick succession (>175ms interval), triggering an audial stimulus and recording the impact data to a file. Hence, great effort has been expended on developing the measurement instrument. This section will detail what steps have been taken in this process, what problems were encountered and how they were solved or circumvented.

3.1 First iteration

The initial measurement instrument design consisted of a dummy of a human torso that contained sensors capable of measuring acceleration. Onto this dummy, images of people were projected with a video projector and participants were asked to hit the dummy with an aluminium baseball bat.

For the first prototype, three analogue ADXL193 acceleration sensors were placed in a triangular configuration in the chest of the dummy. Each sensor is capable of measuring up to 120 g and was sampled at approximately 14 kHz. Upon

1 Rowland, Linehan, Kwamena, & Schoonheyt (2013) have created a punchable device for specific use as a computer interface. Their technology, relying on microphones, is “deemed to approximate the strength of each punch” and is not designed to accurately measure impacts in quick succession.

hitting the prototype with the baseball bat, it however turned out that the measurement range was insufficient and clipping occurred. More importantly, these tests showed that the exact position of the hit influenced the measurement to a great extent. The cumulative measurement, the aggregated vector of all three sensors and the maximum amplitude of each single sensor was analysed. All methods showed great variance dependant on the exact location of the hit.

3.2 Second iteration

To overcome the issue of measuring both the location and the force of the hit, the three acceleration sensors were placed along the X, Y and Z axis in the hollow aluminium baseball bat itself. They were connected via a cable to an enclosure that was hung from the belt or upper arm of the participant. This enclosure housed the electronics, a memory card and a battery, making it a standalone device which could be easily taken to participants. The stimulus remained projected onto a dummy.

Testing a prototype of this setup brought two problems to light. First, the 120 g accelerometers again proved insufficient. To overcome this issue, they were replaced by an ADXL375 three-axis digital accelerometer capable of measuring up to 250 g. However, the baseball bat provided so much momentum that a hit at full swing would easily induce more than 250 times the force of gravity on the sensor. Different positions of the sensor in the bat have been tested and experiments have been conducted with different lengths of the bat. Testing however showed that the impact could not be reliably and accurately measured.

This was partly due to a second problem, which was the maximum sampling frequency of the new, digitally interfaced, sensor. As the moment of peak acceleration, or rather the peak deceleration when hitting the dummy, is very brief, sampling the sensor at the maximum speed of 3200 Hz turned out to be insufficient. Controlled test drops from various heights and on various surfaces, expected to yield internally consistent peak acceleration values, showed that peaks were often not registered. While this was more true for harder surfaces (a laminate floor and a floor mat) than for softer surfaces (a firm mattress and a down pillow), the variance was still unacceptable in the latter conditions. Drop tests have also been performed on the 6 cm thick polyether foam that was sourced for the construction of the dummy, but showed similar inaccuracies in the measurements.

Figure 1 shows an example of the outcome of a controlled drop test to assess the suitability of the ADXL375 sensor for this purpose. Note that in Figure 1 and

subsequent figures, time is very closely approximated by the number of samples on the horizontal axis. As the sampling frequency varies slightly during the measurements, the output is however not exactly proportional to time.

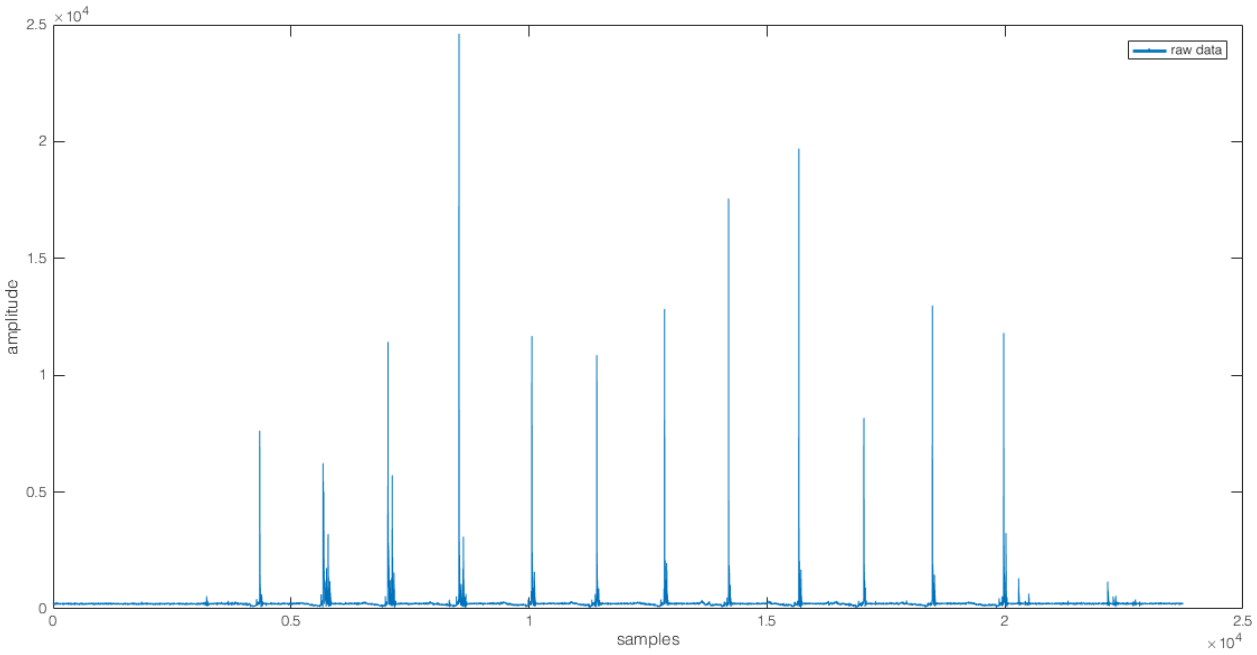

Figure 1: Twelve controlled drops with the ADXL375 in the baseball bat from 16 cm height on a wooden table.

3.3 Third iteration

As the moment of peak acceleration is too brief to measure with the previous designs, an instrument was designed to slow down this movement sufficiently for the sensor to capture it reliably. Taking inspiration from the *makiwara*, a martial arts training device, we built a wooden construction that could be strapped to a post or tree (Figure 2). This device has firm compression springs to provide the resistance necessary to slow down the punch. Initially one compression spring was used, but this proved insufficient. Hard punched would compress the spring fully, which caused a rapid deceleration when hitting the stop and thus clipping the signal. To solve this problem, a second spring was added. This had the added benefit of reducing resonance in the arm after a punch, which causes noise in the signal. A second important feature of the design is the hinge at the bottom. This limits the motion to one direction: back and forth. This means that one ADXL193 acceleration sensor can be used, as opposed to one for each spatial dimension which was

necessary earlier. This allows for a three times higher sampling rate with the same microcontroller. The sensor is positioned along the moving arm of the device and a few sample runs were performed to find the best position for this sensor. A higher position means more movement and thus a better resolution, but also increases the chance of clipping. The device could be strapped to a tree or pillar and rubber backing was added to prevent slipping.

Figure 2. Different views of the 3D-model of the third/fourth design iteration.

The gendered stimulus in the initial experiment setup was a projection onto a human shaped dummy clad in white fabric. However, this had to be changed to a different kind of stimulus as the new device did not allow for a projection. Furthermore, it was found more convenient to take the device to the subjects than the other way around, and video projectors using mains power were deemed not sufficiently mobile. Therefore we switched from a visual stimulus to an audial stimulus, which was triggered by the device. The stimulus is specific to the case study and is therefore further detailed in that section.

3.4 Fourth iteration

The baseball bat was eventually discarded after it was found that it still gave issues with the signal going out of range when the subjects hit more than 120 g. Furthermore, the device had to be mended and reinforced after a participant had broken it. The input from the participants was thus changed to hitting the device

with their bare fists. A major drawback, and the reason to opt for the baseball bat initially, is that the risk of injury is much greater when punching with bare hands. Therefore, a consent form was introduced stating that the participant had no known injuries to hands, wrists, arms or shoulders, nor had reason to believe these were likely to be sustained during the experiment.

A second issue with punching with bare fists was the fact that the minimum time between two hits dramatically decreased. Most participants used two hands and the absence of mass, compared to when swinging the bat, meant they could hit the device even quicker. This posed additional challenges in processing the data, but did lead to more data points per participant and therefore a better estimation of the mean punching force. Apart from switching from a baseball bat to fists, no changes were made to the device.

3.5 Fifth iteration

The device performed well for the first twenty-eight experiments of the case study, but the wooden construction unfortunately proved not strong enough to withstand the most enthusiastic of participants. Some participants tended to punch slightly to the side of the punch cushion, alternating between jab punches and hook punches. Contrary to the jab punches, pushing the cushion backwards, the hook punches also introduced torsional stresses. The long and narrow wooden arm of the device meant there was substantial leverage in the construction, eventually leading to it breaking during one of the trials.

This opportunity was taken to implement the lessons learned from the last design and improve the construction in a final iteration (Figure 3). To improve the lateral stability, the springs were placed in a horizontal instead of a vertical configuration. Again two springs were used, as tests with one spring proved it provided too little resistance and resulted in clipping sensor readings. The wooden arm was discarded and the reinforced cushion was directly hinged from the backplate. The springs were placed at the very top of the backplate. The advantages of this construction were threefold. First, the overall rigidity of the design was improved by eliminating the arm, shortening the device and distributing the lateral forces better over the metal springs. Second, the device was more compact allowing for easier transportation to testing locations. Finally, the lack of mass above the springs greatly reduced unwanted resonance after a hit. Vibrations were absorbed quicker, resulting in cleaner figures with easier to distinguish peaks.

Figure 3. Different views of the 3D-model of the fifth design iteration.

3.5 Complementary technology

The ADXL193 (model AD22282, 120 g) acceleration sensor was connected to an Arduino Due microcontroller, running at 84 MHz. An SD card was inserted into an SD card module which was connected to the Arduino via SPI. A rotary encoder, a button with an LED and a backlit display made sure participant and condition numbers could be set easily. This interface was deliberately designed to look advanced, to convey the message that this was a serious scientific project, despite the peculiar assignment the subjects were given. It also provided quick diagnostics by toggling the LED for each hit and summarising the results of the trial on the display when it was done.

To play the audial stimulus, the microcontroller was connected to a laptop via USB. It communicated with a small Java application that played the sound stimuli. The participants were asked to wear a pair of Nokia BH-604 headphones, that were in turn connected to the laptop via Bluetooth. The A2DP audio profile was used to play the samples in high fidelity, at slightly louder than speaking volume on both channels. We opted for Bluetooth headphones over wired ones to prevent cables from interfering with the movement of the subject. A second benefit of using headphones instead of speakers was that the stimuli were not audible for bystanders, which turned out to be vital given the nature of the samples used in the case study.

4. Case study

To test the method, we performed a small case study where we applied the newly devised research method to the concepts of aggression and gender. This not meant as validation, but merely to see whether meaningful results can be obtained with this method. The outcome should therefore be approached with caution.

4.1 Research question and hypothesis

In this case study, we sought to find *whether genderisation of an inanimate object modulates the physical aggression exhibited towards it*. Given the previous sections, one could expect a more ambitious research question, that does not specify the aggressed against as an inanimate object. Taking into account that this study merely serves as a proof-of-concept for the created method for aforementioned reasons, we deemed it inappropriate to make inferences about gender relations in general. In other words, we do not mean to study whether the subjects are more aggressive towards males or females, but merely whether and how gender modulates their aggressive behaviour towards the device. Following this, we formulate the following null-hypothesis:

$H1_O$: The subjects exhibit the same level of aggression towards the device when it is genderised male as when it is genderised female.

This is accompanied by the following alternative hypothesis:

$H1_A$: The subjects exhibit a different level of aggression towards the device when it is genderised male than when it is genderised female.

Eagly and Steffen (1986) state in their meta-analytic review that men tend to temper their aggression towards women, especially in laboratory settings and when the aggression is physical. Eastin (2006), who conducted various experiments involving gendered avatars in a video game, concluded similarly that men tend to exhibit more aggression towards male opponents than towards female opponents. We formulated a second hypothesis in line with these findings:

H2: subjects exhibit more aggression towards the device when it is genderised male than when it is genderised female.

4.2 Research design

As the expected variance in punch impact is large, we opted for a within-subjects design. This allows for a smaller number of test subjects as each subject is compared with itself, yet is prone to the carryover effect and fatigue. To prevent the first from introducing a structural bias, the order of the conditions to which the participants are exposed is altered. This is known as a counterbalanced measures design. To prevent fatigue, the sessions are limited to thirty seconds and there is a break of at least one minute in between the two sessions.

4.3 Sample

The sample considered for analysis consisted of 22 men, aged between 16 and 31 years old (mean = 23.8, SD = 4.3). They were recruited on four non-consecutive days in a park in Amsterdam, at two locations of Leiden University, and a park in Leiden. Data was gathered for 42 participants, but much of it was deemed not fit for analysis. Due to technical problems, the data of the first ten participants was recorded at an insufficient rate of 4.8 kHz. Seven more participants were excluded as they failed to hit the device hard enough in one or both of the trials to trigger the stimulus[2]. One participant had to be excluded because the device broke during the measurements. Another subject's data was removed from analysis for he was deliberately "trying to manipulating the data", as he literally phrased it during the experiment. Finally, when the sample consisted of two women and twelve men, the two women were removed from analysis. During the gathering of the first data, we found that both men and women were initially somewhat hesitant to participate. However, men seemed easily persuaded by the remark that their contribution was of a physical nature, while women seemed deterred by that. Following this insight, the research was scoped on men and ten more male participants were drafted.

4.4 Stimulus

The sound stimulus consisted of grunts and ouches; sounds that someone would typically make when they are punched. To prevent repetition which would be perceived as fake, ten sound samples were sourced for each gender and were played

2 The stimulus was triggered at an impact of approximately 2.9 g (device iteration IV) and 1.4 g (device iteration V). This is roughly comparable to a manly pat on the back, but would not qualify to most as a punch.

sequentially. The samples were freely available on the internet[3]. The stimulus was triggered with a minimum interval of one second, as this came across as more natural to the researchers.

It should be noted that the sounds were not validated before the experiment. It can be argued that other connotations carried by voice besides gender influences the punches. The degree of pain that is expressed could also influence the participants' behaviour. It is good practice to first establish whether the sound samples are equal by subjecting people of both genders to them and having them score the sounds in various measures for pain and suffering. A potential correction factor can then be applied. However, due to constraints in time and means, this was out of scope for this study.

4.5 Protocol and data gathering

The participants were asked for their first name, their gender and their age. No other information that could identify them was asked, to ensure their anonymity. After they had signed the waiver, a short explanation of the experiment was given. The participants were told that the experiment encompasses two sessions and that they got to wear a pair of headphones (Figure 4). Each session lasted thirty seconds and there was a one-minute break in between the sessions. Finally, it is explicitly made clear that they can hit the device as hard and as often as they feel like. When participants asked whether they should hit the device hard, often, or with one or two hands, the same instruction was repeated. A researcher timed the sessions with a stopwatch and told the participants when to start and when to stop. The thirty seconds sessions started on the first punch. The device recorded for forty seconds, to provide a little margin for participants starting late.

---

3 The female samples were downloaded from https://www.freesound.org/people/AderuMoro/sounds/213295/ on 20 June 2016. The male samples were downloaded from multiple pages on Soundbible.com on the same day, but the website is offline at the time of writing. The selected files can be downloaded from http://jeroenvanoorschot.nl/stimulus.zip

Figure 4. A participant wearing headphones and punching device iteration V during the experiment, with the control box on the right side and a laptop in the backpack.

As two of the testing locations were the Vondelpark and the Van der Werfpark, two public parks in Amsterdam and Leiden respectively, onlookers were sometimes present. However, due to the participant wearing headphones, onlookers were not aware of the stimulus nor the order. Especially in the Vondelpark, sports such as frisbee, tight-rope walking, running, skating and bootcamp are common, so the activity was deemed not too much out of place.

As detailed in the iterative method design section, the fourth iteration of the device got severely damaged during data gathering and was replaced. In the sections below, we use *iteration IV* to refer to the device used with the first twelve participants (Figure 2). We refer to the device used with the last ten participants as *iteration V* (Figure 3).

The mean amplitude reported for participants using iteration V is lower (226.9) than for those using iteration IV (261.6). This is due to the absence of the long arm, which provided more leverage and effectively lowering the resistance of the springs in the earlier design. However, the within-subject research design that is used is resilient to these changes. The characteristics of the sensor, the stimulus nor the basic measurement principles changed, and both conditions for each subject were measured on the same device. Hence, we have no reason to assume that the

changes in the measurement instrument after gathering data for the first half of the participants introduced a systematic bias.

4.6 Data processing

The microcontroller recorded approximately 1.6 million to 1.75 million[4] data points per forty second recording session. This translates to a sampling frequency of approximately 40 kHz to 42.5 kHz. As the microcontroller measured with a twelve-bit resolution, the values returned were between 0 and 4095 inclusive. The generated data files were imported into MATLAB, where the mean value in idle (2009) was subtracted from all measurements to centre the values around zero. Of interest are the negative peaks, or valleys, as they represent backward acceleration of the sensor caused by the participant's fist hitting the device. The positive peaks represent the forward acceleration, caused by the springs pushing the arm back to its original position. Peaks are also found when the arm is punched back beyond the point where the springs are fully compressed. This occurred only in two cases, resulting in values clipping at the maximum sensor value. However, this did not affect the measurement as the clipping occurred in the positive domain.

Initially a peak detection algorithm (Yoder, 2015) was applied to this data, in order to retrieve the amplitude and location of the punches. However, the nature of the data caused the algorithm to not perform as well as desired (Figure 5a & 5b). This can partly be attributed to the aforementioned problem of people hitting the device with both hands in quick succession. This resulted in a large number of difficultly distinguishable low peaks, instead of a smaller number of easily distinguishable high peaks as was the case with the baseball bat. The fastest puncher recorded on average over 3.8 hits per second.

As a bias is easily introduced by applying powerful algorithms to noisy data (Lamers, 2015), extra caution was taken to prevent this from happening. The optimal selectivity parameter of the peak detection algorithm was defined as the number that would make sure the algorithm would return all visible peaks and the lowest number of false positives. In a preliminary analysis of five participants, this parameter was determined per participant for both conditions. This proved that the optimal selectivity parameter was vastly different for each participant and condition,

4 The difference is explained by the use of different memory cards. Peculiarly, the faster Sandisk Extreme Plus 16Gb 45 MB/s card achieved a lower sampling rate (~40 kHz) than the Kingston 2Gb 1.5 MB/s card (~42.5 kHz).

which troubles both within and in between subject comparison. Furthermore, the amount of erroneously identified peaks, which had to be filtered out by the experimenter based on visual analysis of the plots, was over 83% in some cases.

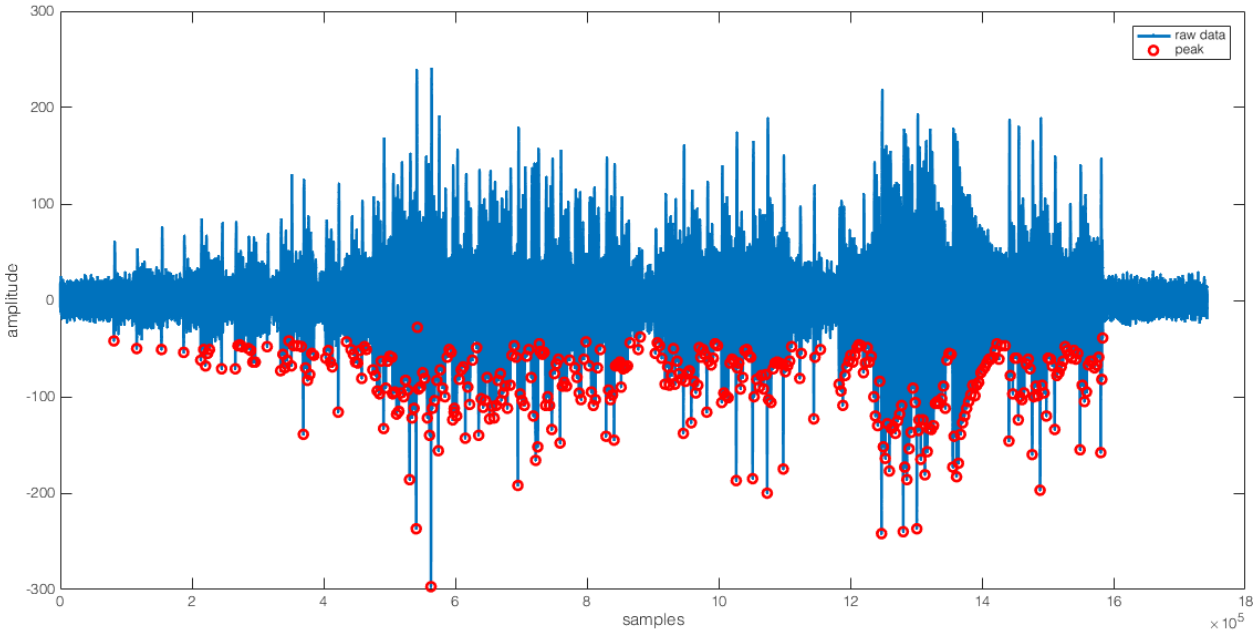

Figure 5a: Example plot of a single trial using device iteration IV after applying the peak detection algorithm to the raw data: 376 peaks are returned.

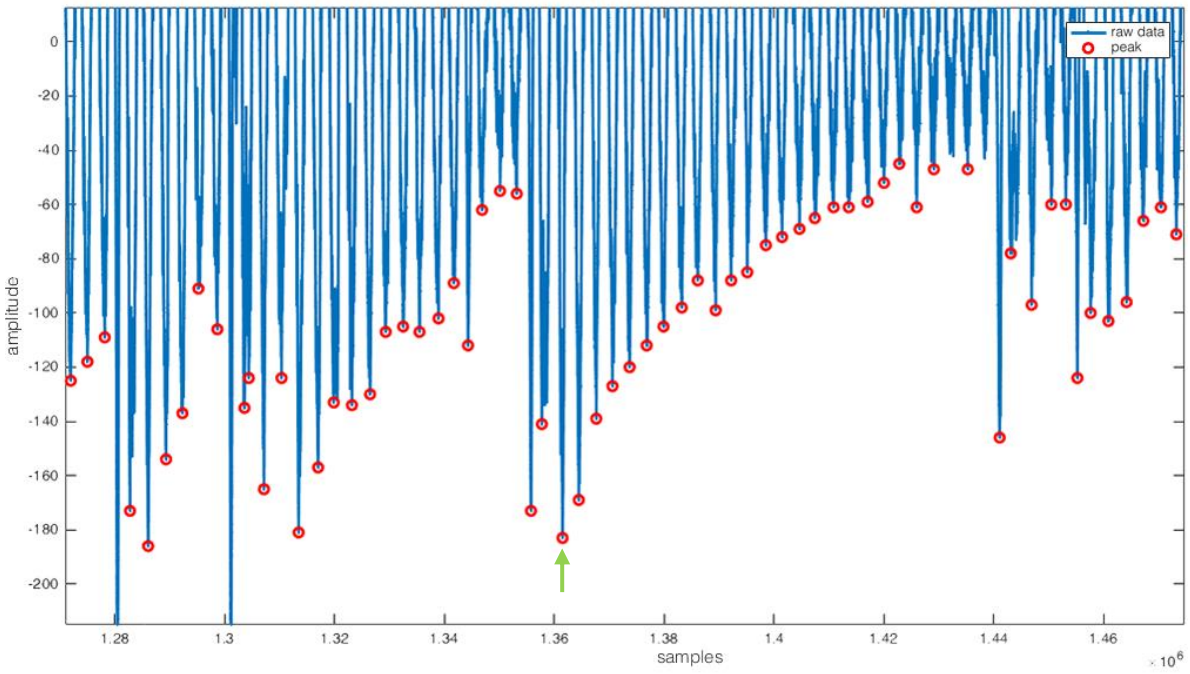

Figure 5b: Close-up of Figure 5a. The decay of the marked hit is registered as separate hits and numerous other false positives are returned.

To overcome this problem, we wrote a two-stage algorithm to transform the data before subjecting it to the peak detection algorithm. First, we applied a Savitzky-Golay finite impulse response (FIR) smoothing filter to remove some of the noise. The benefit of this type of filter is that it preserves the high frequency components better than regular FIR filters. Due to the nature of the data, the duration of the peaks is very short. The Savitzky-Golay filter is based on polynomial approximation of local least-squares and preserves the peak values that are of interest to us better (Orfanidis, 1996). To minimise the undesired effects of the filter, a small polynomial order (3) and an extremely short data frame length (5) are chosen. While this step removed only a few false positives if the peak detection algorithm was applied to the filtered data (Figure 6), it proved sufficient to successfully perform the next step.

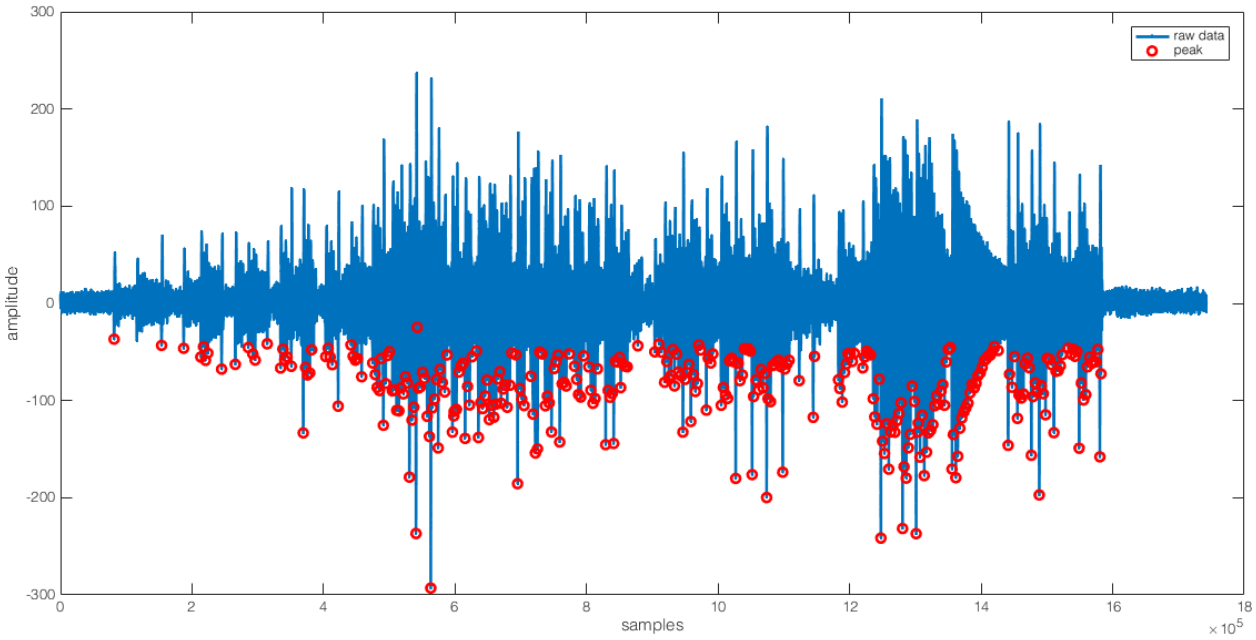

Figure 6: Example plot of a single trial using device iteration IV after applying the peak detection algorithm to the filtered data: 331 peaks are returned.

In the second stage, we determined the upper and lower peak envelope of the signal. The was done with the MATLAB *envelope* function, that uses spline interpolation over local maxima to draw a smooth curve along the peaks (MathWorks, 2016). We set the minimum peak separation to 3500 samples, which translates to approximately 80 ms at a sampling rate of 40 kHz. The previously

mentioned preliminary manual analysis showed that valid punches were spaced at least 200 ms apart, so this value was deemed sufficient. Lowering the minimum peak separation further increased the tendency to incorrectly assign peaks to non-peak data. Subsequently, the peak detection algorithm was applied to the lower signal envelope (Figure 7). This yielded results in accordance with the number of punches as roughly estimated by the researcher during the trials.

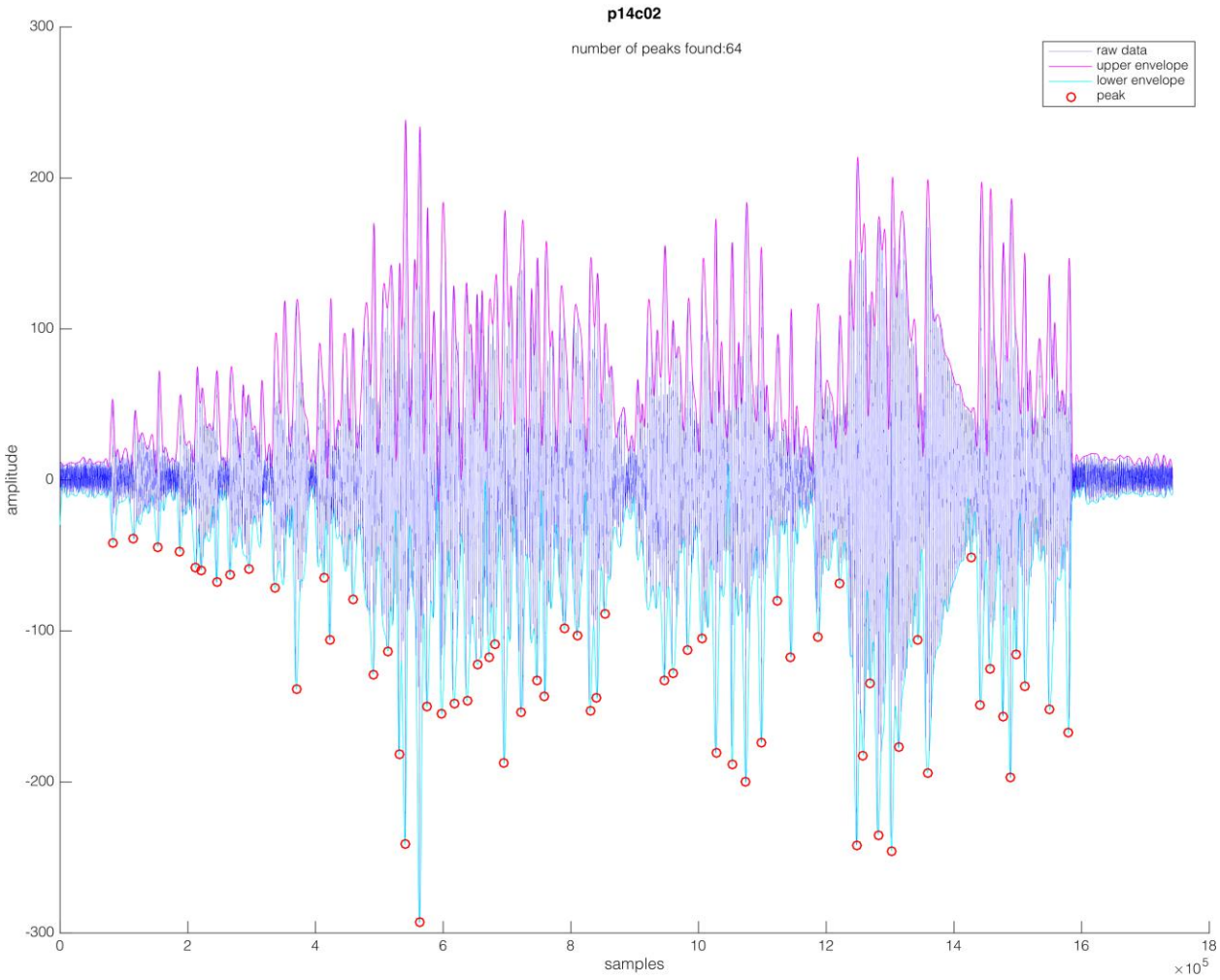

Figure 7: Example plot of a single trial using device iteration IV after applying the peak detection algorithm to the lower signal envelope: 64 peaks are returned.

The interpolative quality of the signal envelope caused some of the peaks to be marked above the raw data. No peaks have been found for which the amplitude returned by the algorithm is > 30 higher than the raw data reflects. Furthermore,

this error appears to affect all trial data consistently. We therefore assume that this procedure introduces no systematic error.

The iteration IV device caused the algorithm to still return a very small number of false positives, 2.6 per trial on average (SD = 3.4). These were removed from analysis by the experimenter based on visual inspection of the plotted raw data and signal envelope. It was suspected these occur because of noise caused by the resonating arm of the iteration IV device after a punch. Indeed, the trials with iteration V showed less noise and no false positives had to be removed (Figure 8).

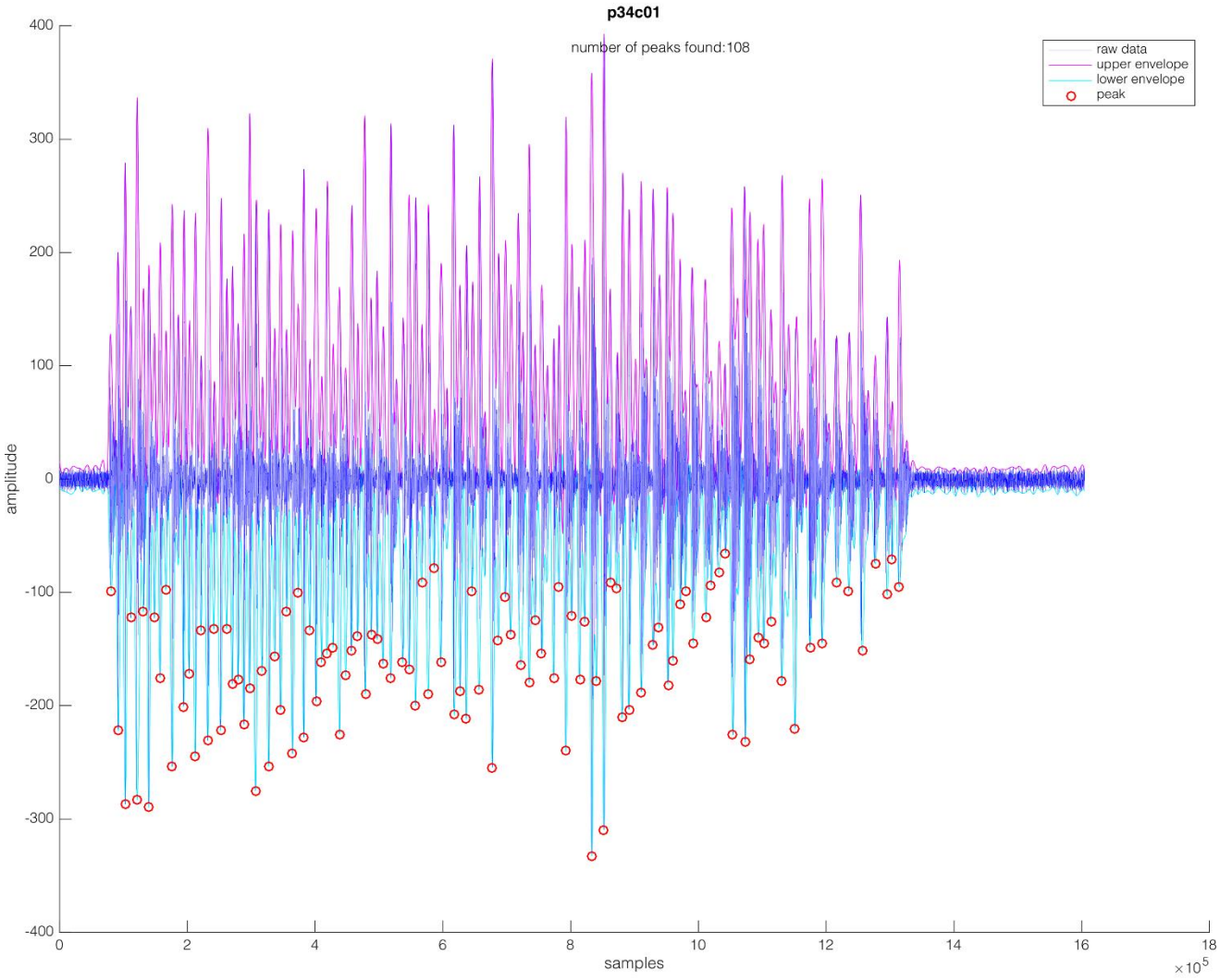

Figure 8: Example plot of a single trial using device iteration V after applying the peak detection algorithm to the lower signal envelope.

For methodological reasons, all data up to and including the third punch that triggered the stimulus were removed from analysis. During the experiments, subjects remarked that it took some time before they fully recognised the stimuli as being male or female. We discarded all hits before the threshold value was reached three times, reasoning that these hits were not guaranteed to be subject to the stimulus. For design iteration IV, the threshold value was set at -100 amplitude, as a more tolerant value would trigger the stimulus repeatedly due to resonance. The value was changed to -50 amplitude for iteration V, as less resonance allowed for a lower threshold.

Finally, the data were inverted to the positive domain in order to ease analysis and prevent confusing double negative phrasing. As all peaks were negative and only the distance from zero is of interest to us, this simply meant taking the modulus of all values. Thus, in the remainder of the text, a higher amplitude represents a harder punch.

4.7 Results

All but two participants fully used the thirty seconds they were given. A foreign participant, having the female stimulus in the second trial, declared that in his culture “it is wrong to hit women”. He returned the headphones after ten seconds, having registered six hits versus twenty-five in the male condition. Another participant simply waited for the second half of both sessions to end, commenting that he “didn’t feel like continuing”. Both participants were not excluded from analysis, as they complied with the assignment and registered sufficient hits before stopping.

The peak detection algorithm returned for each participant the peak amplitudes of the individual punches in each condition, marked by the red dots in Figures 7 and 8. These peak values were subjected to a two sample t-test assuming unequal variances, in order to establish whether the punches from the two conditions were statistically different (Table 1). A mean amplitude for these punches per condition per participant is reported, as well has the number of hits. Looking at the individual subjects, we see a significant difference is found between the two conditions for twelve out of twenty-two subjects. Table 1 also provides the order and the iteration of the measurement device for each participant.

			male stimulus			female stimulus			
device	ID	order	number of hits	mean amplitude	SD amplitude	number of hits	mean amplitude	SD amplitude	P-value
IV	1	FM	76	**180.7**	67.4	83	**262.9**	101.3	**0.000**
	2	MF	25	862.8	251.6	6	801.8	214.9	0.561
	3	FM	60	107.1	50.2	42	125.9	57.9	0.093
	4	FM	66	**179.4**	71.0	45	**339.4**	200.5	**0.000**
	5	MF	97	340.3	109.3	112	339.5	100.9	0.955
	6	MF	77	226.1	94.5	61	226.8	108.9	0.972
	7	FM	84	**338.8**	110.8	91	**276.5**	90.7	**0.000**
	8	MF	94	**83.0**	30.7	91	**73.3**	26.1	**0.022**
	9	FM	76	**124.7**	41.7	81	**144.7**	46.6	**0.005**
	10	MF	81	**157.3**	56.7	72	**225.6**	90.6	**0.000**
	11	FM	60	98.9	38.1	57	101.3	53.5	0.782
	12	MF	7	381.9	45.3	4	279.8	73.9	0.066
V	13	FM	88	**368.8**	169.1	89	**285.9**	102.7	**0.000**
	14	FM	100	357.7	167.8	41	334.7	164.3	0.455
	15	MF	105	**164.0**	55.2	100	**91.6**	33.1	**0.000**
	16	FM	70	224.3	87.1	81	213.6	83.4	0.443
	17	MF	45	**225.3**	88.1	55	**153.1**	66.3	**0.000**
	18	FM	96	250.2	82.4	85	267.0	80.6	0.167
	19	MF	50	**249.9**	155.0	62	**186.0**	92.1	**0.012**
	20	FM	77	**108.7**	65.8	83	**314.3**	189.4	**0.000**
	21	FM	49	265.4	142.6	74	221.7	102.9	0.068
	22	MF	63	**157.9**	92.1	76	**97.4**	50.7	**0.000**

Table 1: Summary of a two-sample t-test per subject assuming unequal variances comparing mean amplitudes. Two-tailed significance is reported.

While we tried to eliminate order effect by implementing a counterbalanced measures design, we also performed a post-check for all subjects that reported a significant difference between the two conditions. We calculated the difference between the mean amplitudes for each condition, by subtracting for each participant the mean amplitude for the female condition from the mean amplitude for the male condition (as reported in Table 1). This difference is referred to in Table 2 as *raw Δ mean amplitude*. We performed a two-sample t-test assuming unequal variances on the raw Δ mean amplitudes of the two different groups: male condition first and female condition first. This indicated that the difference in mean amplitudes is not statistically different for the raw amplitudes ($p > .137$). Following the same procedure, the *standardised Δ mean amplitudes* have been calculated based on the standardised mean amplitudes (as reported in Table 3). These were subjected to the same t-test, which also returned an insignificant difference ($p > .075$). This indicates that the reported mean amplitudes are not influenced by the order in which the stimuli are experienced. However, it should be noted that the number of participants qualifying for this post-check is extremely small and the *p*-value found for the test on the standardised Δ mean amplitudes is low. Order effects should therefore be investigated thoroughly in future studies.

1st: female 2nd: male			1st: male 2nd: female		
ID	raw Δ mean amplitude	standardised Δ mean amplitude	ID	raw Δ mean amplitude	standardised Δ mean amplitude
1	-82.2	-0.858	8	9.7	0.335
4	-160.0	-1.006	10	-68.3	-0.834
7	62.3	0.592	15	72.4	1.242
9	-20.0	-0.442	17	72.3	0.855
13	83.0	0.571	19	64.0	0.501
20	-205.6	-1.164	22	60.4	0.772

Table 2: Difference in mean amplitudes, raw and standardised, between the two conditions, grouped by condition order.

To answer the research question in a statistically satisfying manner, we need to analyse the sample as a whole. As can be seen in Table 1, the mean amplitude varied much between the individual participants, which does not allow for straight comparison of the individual amplitudes between the subjects. To facilitate this, we need to normalise the effect of individual characteristics such as strength on the recorded amplitudes.

We standardised the amplitudes ($\hat{a}$) of each participant (p), which allows us to express the amplitude in multiples of its standard deviation above mean. The following formulas stress that population mean μ and population standard deviation σ for each participant were estimated by the mean and standard deviation of all hits recorded for that participant. By using the mean and standard deviation of the amplitudes from both conditions aggregated and not for each condition separately, we calculated the μ and σ for the participant instead of for the different conditions. This allows for comparison between the two conditions.

$$\hat{a}_{p_{male}} = \frac{a_{p_{male}} - \mu_p}{\sigma_p} \qquad \hat{a}_{female} = \frac{a_{female} - \mu}{\sigma}$$

As standardisation is merely translation followed by scaling, we expected the result of a t-test on the amplitudes for each participant to be equal to Table 1. To check whether this supposition is correct and the standardisation did not discard information we wish to preserve, we performed a t-test for each participant on the standard scores. As expected, the same p-values are found for the same participants (Table 3).

			male stimulus			female stimulus			
device	ID	order	number of hits	standardised mean amplitude	SD standardised amplitude	number of hits	standardised mean amplitude	SD standardised amplitude	p
IV	1	FM	76	**-0.448**	0.495	83	**0.410**	1.057	**0.000**
	2	MF	25	0.049	1.074	6	-0.203	0.885	0.561
	3	FM	60	-0.143	0.863	42	0.205	1.072	0.093
	4	FM	66	**-0.408**	0.200	45	**0.598**	1.261	**0.000**
	5	MF	97	0.004	1.091	112	-0.004	0.965	0.955
	6	MF	77	-0.003	0.880	61	0.003	1.081	0.972
	7	FM	84	**0.308**	1.107	91	**-0.284**	0.862	**0.000**
	8	MF	94	**0.165**	1.131	91	**-0.170**	0.904	**0.022**
	9	FM	76	**-0.228**	0.849	81	**0.214**	1.029	**0.005**
	10	MF	81	**-0.393**	0.480	72	**0.442**	1.107	**0.000**
	11	FM	60	-0.025	0.686	57	0.027	1.161	0.782
	12	MF	7	0.499	0.372	4	-0.874	0.995	0.066
V	13	FM	88	**0.287**	1.353	89	**-0.284**	0.499	**0.000**
	14	FM	100	0.040	1.015	41	-0.098	0.973	0.455
	15	MF	105	**0.606**	0.896	100	**-0.636**	0.322	**0.000**
	16	FM	70	0.068	1.050	81	-0.058	0.962	0.443
	17	MF	45	**0.470**	1.084	55	**-0.385**	0.615	**0.000**
	18	FM	96	-0.097	1.016	85	0.109	0.971	0.167
	19	MF	50	**0.277**	1.475	62	**-0.224**	0.521	**0.012**
	20	FM	77	**-0.604**	0.139	83	**0.560**	1.150	**0.000**
	21	FM	49	0.216	1.375	74	-0.143	0.715	0.068
	22	MF	63	**0.422**	1.386	76	**-0.350**	0.420	**0.000**

Table 3: Summary of a two-sample t-test per subject assuming unequal variances comparing mean amplitudes after standardisation. Two-tailed significance is reported.

The standardised amplitudes of all subjects under each condition were aggregated and subjected to a two-sample t-test assuming equal variances[5] (Table 4). This was first done for participant 1 through 12 as a preliminary analysis of the data from device iteration IV. This indicated that the standardised mean amplitude was higher under the female condition than under the male condition. However, a separate analysis of the data gathered with device iteration V for participant 13 through 22 showed a statistically significant effect in the opposite direction. These effects cancel each other out and therefore no significant effect is found for the whole sample.

		male stimulus			female stimulus			
device	participants	aggregated number of hits	standardised mean amplitude	SD standardised amplitude	aggregated number of hits	standardised mean amplitude	SD standardised amplitude	p
IV	1 - 12	803	**-0.092**	0.927	745	**0.099**	1.058	**0.000**
V	13 - 22	743	**0.153**	1.069	746	**-0.153**	0.895	**0.000**
IV & V	1 - 22	1546	0.026	1.005	1491	-0.027	0.987	0.144

Table 4: Two-sample t-tests for data gathered with iteration IV, iteration V, and iteration IV & V combined, assuming equal variances comparing aggregated standardised amplitudes for both conditions.

[5] Unequal variances were assumed for the raw scores in earlier tests. However, standardisation resulted in a fixed standard deviation of 1 for each participant. We therefore assume the variance to be equal.

4.8 Analysis

Following these results, we accept hypothesis $H1_0$, stating that the subjects exhibit the same level of aggression towards the inanimate object when it is genderised male as when it is genderised female. Consequently, both the alternative hypothesis $H1_A$ and hypothesis H2, suggesting the direction of the effect, are rejected.

While the stimulus, the instructions, the electronics and the method did not change when device iteration V was used, the differences between the data gathered with the two devices as shown Table 4 is worthy of investigation. We therefore assumed a similar approach as we did for controlling for order effects (Table 2). We performed a two sample t-test assuming unequal variances on the mean amplitudes of both conditions registered for each device. When comparing the raw mean amplitudes in Table 1 between iteration IV and iteration V, we found no significant difference ($p > .442$). An even less substantial difference was found between the two devices when comparing the standardised mean amplitudes from Table 3 ($p > .855$). This suggests that the different effects found for trials with different devices, as described in Table 4, did not influence the results. Assuming no structural differences in the samples, we can only conclude that the opposite effects found with iteration IV and V is spurious. The small sample sizes of twelve and ten men respectively easily introduces such effects.

Looking at the individual participants, we find that ten do not report a significant difference between the two conditions. It can be argued that they experienced the device not as sufficiently male or female, or even sufficiently human. The stimulus therefore did not modulate the punching force consistently. Studies in video game research have shown that moral disengagement cues, such as the non-anthropomorphic appearance of the device, frame violence as acceptable (Hartman & Vorderer, 2010). As the audial stimuli are a synecdoche for gender and humanness at best, it is possible that insufficient characteristics are conveyed to trigger behaviour in line with the subjects' social norms.

For twelve participants, a significant difference between the two conditions is found. The anecdote of the subject stopping halfway the female condition, stating that in his culture "it is wrong to hit women", suggests that he perceived the device as intrinsically female. The degree to which the participants perceived the device as

male, female, or human in general was beyond the scope of this case study, but is likely to be a confounding factor.

Out of the twelve significant cases, seven participants report a higher mean amplitude under the male condition, while five report a higher mean amplitude under the female condition (Table 1). This does not show a general trend, nor is it the aim of this case study to infer the underlying reasons for the differences found in the individual subjects.

However, the five cases where a significantly higher mean amplitude was reported for the female condition are noteworthy. Eagly & Steffen state that aggression research involves “some type of explicit or implicit surveillance of subjects' behavior, which should often heighten the salience of norms that temper men's aggression toward women” (1986, pp. 312-313). Considering this, it is all the more remarkable that 23% of all participants did not temper their aggression in the female condition. This number increases to 42% if we only consider the participants for which we assume that the stimulus affected how much aggression they exhibited. It could be argued that for these participants, the use of a physical, non-cognitive research method did not heighten the salience of cognitive norms. In other words, the method might inhibit the cognitive social desirability bias to intervene with physical expression of aggression in certain cases.

5. Discussion

The case study has shown that even with a small sample, interesting findings can be obtained. That being said, it is not without its difficulties. In this section we will detail some of the key shortcomings of the method and how they affected the outcome.

We encountered numerous difficulties in conveying the notion of gender. Two participants reported that they had trouble recognising the female audio stimulus as a painful female outcry. One remarked that some of the grunts came across as if the woman was delivering a punch instead of receiving it. Another said that it sounded like a child to him. The whole exercise reminded him of bit of a computer game such as Street Fighter for the Super Nintendo. We think that the gender stimulus can be expanded upon in future studies. One could play a short introduction in the voices of the male and female stimuli, sharing some personal details with the participant in order to build rapport. The stimulus can be further extended upon by introducing visual and even olfactory gender cues.

We should also note that response bias is possibly not completely mitigated, as the respondents knew that the researcher would analyse their data and was watching during the trials. This bias is arguably also introduced when respondents analyse the goals of the study during the trials. A difference was perceived between respondents drafted at the university and random respondents drafted on the streets. The first were much more inquisitive during the one minute break between the two trials and after the experiment. However, factoring in education levels was beyond the scope of this study. A more covert strategy, for example disguising the measurement rounds as practice rounds, can possibly mitigate this. This was considered, but eventually dismissed for practical reasons.

The method is difficult to generalise from a genderised object to a gender as such, without substantial theoretical backing and validation through other methods. However, we do believe that embedding a similar method in a research design that is more extensive than the case study we presented, allows for validation of the findings through cross verification with other measures.

In addition to these methodological difficulties, there are technical choices in the design of the instrument and the method that affected the outcome. For example, we used the peak measurement instead of the surface under the peaks as a measure for aggressiveness. There might be a difference between people punching fast with less mass and people punching slow with more mass. The latter would create a lower and wider peak with the same kinetic energy, which is not accounted for now. Yet, it can be argued that, assuming equal kinetic energy, a fast jab is more damaging than a slow blow. This is due to the deformity of the target's body caused by the initial transfer of energy, which is higher for faster punches. Considering this, we argue that the peak values are the preferred measurement for aggression.

Finally, in iteration IV, the lack of a shock absorber in conjunction to the springs generated some noise due to resonance. This noise itself did not influence the measurments as such, as it was successfully filtered out by the algorithms. However, it is possible that punching the cushion while it is still resonating from the last punch influences the measurement of the next punch. This is especially true for punches in very quick succession. Device iteration V was less affected by distortive resonance, as the sustain was much shorter. The sensor had stopped moving before the next punch was delivered.

6. Reflection

Scientists go to great lengths to counter the effects of social desirability bias in aggression research, and the ingenuity of their efforts is worthy of praise (see Saunders (1991) for some examples). Considering this, our method is not a replacement for questionnaires, interviews and other self-report methods — nor does it intend to be. Self-report methods have proven extremely useful in gathering data on personality characteristics, feelings, preferences, opinions, et cetera. However, due to this success, we feel that oftentimes alternatives are prematurely dismissed, or not considered altogether. This paper is also an attempt to show that alternative methods, even fairly low-tech ones such as the method proposed in this study, are worthy of exploration.

We state from personal experience that students in the social sciences are trained in perfectly executing the methods that have been around for a long time. They are skilled in applying a plethora of measures and statistical tools to compensate for their shortcomings. Yet, they are not trained to come up with new methods to tackle old and new problems. As technology progresses, it is often offered as just another tool to assist in existing methodology. The technology used in this paper is capable of measuring with extremely high resolution and accuracy, and can be built for less than fifty euros. Research disciplines such as biology and physics have already adopted powerful, low-cost technology to perform experiments that are valuable to science (see for example Sheinina, Lavi, & Michaelevski (2015) and openPCR.org). We believe that these technologies unlock endless possibilities to research problems in creative ways, in all scientific disciplines. They just have not hit the social sciences yet.

Works Cited

Archer, J., Holloway, R., & McLoughlin, K. (1995). Self-Reported Physical Aggression Among Young Men. *Aggressive Behavior, 21*(5), 325-342.

Buss, A. H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology, 63*(3), 452-459.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in oganizational behavior research. *Journal of Business and Psychology, 17*(2), 245-260.

Eagly, A. H., & Steffen, V. J. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100*(3), 309-330.

Eastin, M. S. (2006). Video Game Violence and the Female Game Player: Self- and Opponent Gender Effects on Presence and Aggressive Thoughts. *Human Communication Research, 32*(3), 351-372.

Fiske, D. W., & Pearson, P. H. (1970). Theory and techniques of personality measurement. *Annual Review of Psychology, 21*, 49-86.

Gregoski, M., Malone, W. A., & South Richardson, D. (2005). Measuring direct and indirect aggression: is there a response bias? *Psychological Reports, 97*, 563-566.

Harris, M. B., & Knight-Bohnhoff, K. (1996). Gender and Aggression I: Perceptions of Aggression. *Sex Roles, 35*(1), 27-42.

Hartman, T., & Vorderer, P. (2010). It's Okay to Shoot a Character: Moral Disengagement in Violent Video Games. *Journal of Communication, 60*(1), 94-119.

Hult, G. T. (2011). Addressing Common Method Variance: Guidelines for Survey Research on Information Technology, Operations, and Supply Chain Management. *Transactions of Engineering management, 58*(3), 578-588.

Johnson, N., Carran, S., Botner, J., Fontaine, K., Laxague, N., Nuetzel, P., . . . Tivnan, B. (2011). Pattern in Escalations in Insurgent and Terrorist Activity. *Science, 333*(1068), 81-84.

Lamers, M. H. (2015, March 3). *Artificial neurons & Deep learning*. Lecture at Leiden University, Leiden.

Leyton, E. (2003). Agression: Evolutionairy and Anthropological Theories. In E. Hickey, *Encyclopedia of Murder and Violent Crime* (pp. 8-11). Thousand Oaks: Sage Publications.

MathWorks. (2016, February 22). *envelope*. Retrieved August 18, 2016, from MathWorks Documentation: http://nl.mathworks.com/help/signal/ref/envelope.html

Morren, M., & Meesters, C. (2002). Validation of the Dutch Version of the Aggression Questionnaire in Adolescent Male Offenders. *Aggressive Behavior, 28*(2), 87-96.

Orfanidis, S. J. (1996). *Introduction to Signal Processing*. Upper Saddle River, N.J.: Prentice Hall.

Robins, R. W., Tracy, J. L., & Sherman, J. W. (2007). What Kinds of Methods Do Personality Psychologists Use?: A Survey of Journal Editors and Editorial Board Members. In R. Robins, R. Fraley, & R. Krueger, *Handbook of Research Methods in Personality Psychology* (pp. 673-678). New York: The Guilford Press.

Rowland, D., Linehan, C., Kwamena, A., & Schoonheyt, M. (2013). Return of the Man-Machine Interface: Violent Interactions. In D. Reidsema, H. Katayose, & A. Nijholt (Ed.), *10th International Conference on Advances in Computer Entertainment*. *8253*, pp. 215-229. New York: Springer-Verlag.

Saunders, D. G. (1991). Procedures for Adjusting Self-Reports of Violence for Social Desirabiliy Bias. *Journal of Interpersonal Violence, 6*(3), 336-344.

Sheinina, A., Lavi, A., & Michaelevski, I. (2015). StimDuino: An Arduino-based electrophysiological stimulus isolator. *Journal of Neuroscience Methods, 243*, 8-17.

Vigil-Colet, A., Ruiz-Pamies, M., Anguiano-Carrasco, C., & Lorenzo-Seva, U. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema, 24*(2), 310-315.

Yoder, N. (2015, December 14). *Peakfinder*, version 2.0.1. Retrieved June 22, 2016, from MathWorks File Exchange: https://nl.mathworks.com/matlabcentral/fileexchange/25500-peakfinder-x0--sel--thresh--extrema--includeendpoints--interpolate-