Consciousness and Cognition 65 (2018) 1–26

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Libet's experiment: A complex replication

Tomáš Dominik[a,*], Daniel Dostál[a], Martin Zielina[b], Jan Šmahaj[a], Zuzana Sedláčková[a], Roman Procházka[a]

[a] Department of Psychology, Palacký University Olomouc, Křížkovského 511/8, Olomouc, Czech Republic
[b] Department of Medical Ethics and Humanities, 2nd Faculty of Medicine, Charles University in Prague, V Úvalu 84, Prague 5, Czech Republic

ARTICLE INFO

Keywords:
Libet's experiment
Replication
Rotating-spot method
Subjective event timing
Readiness-potential

ABSTRACT

Libet's experiment is an influential classical study, which does not stop provoking heated debates. However, a full-scale replication has not been carried out to this day. Libet-style studies have usually focused on isolated ideas and concepts and never on the whole experiment in all its complexity. This paper presents detailed methodological description and results of a complex replication study. The methodology follows Libet's directions closely in most cases; when it does not, the differences are described and elaborated. The results replicate Libet's key findings, but substantial differences were found in some of the results' categories, such as the introspective reports or the number of readiness-potentials found. The discussion also addresses some current problems pertaining the methodology of the Libet-style experiments and provides some recommendations based on a detailed process evaluation.

1. Introduction

Occasionally, it happens in science that a thought-provoking empirical study becomes substantially influential and stimulates decades of follow-up debates and research. Such is the case of a study published by Benjamin Libet and his colleagues more than 30 years ago (Libet, 1985; Libet, Gleason, Wright, & Pearl, 1983; Libet, Wright, & Gleason, 1982). Their publication initiated heated debates concerning its methodology and interpretations between other researchers (e.g. Libet, 1985, Open Peer Commentary; Dennett & Kinsbourne, 1992; Breitmeyer, 2002; Klein, 2002; Pockett, 2002; Klemm, 2010; Papanicolaou, 2017) and Libet himself responded to many of the discussions in published papers and his 2004 monograph (e.g. Libet, 1985, Author's Response; Haggard & Libet, 2001; Libet, 2004).

Therefore, it comes as no surprise to learn that some of the technical details of said experiment are, methodologically speaking, not perfect (e.g., see the discussion between Gilberto Gomes and Benjamin Libet in Gomes, 1998; Libet, 2000; Gomes, 2002). Hence, one would expect that there would be a large initiative to replicate Libet's experiment soon after its publication. Nevertheless, the empirical studies replicating Libet's experiment in some way or another seem to adopt a different approach. Vast majority of these studies substantially modified or simplified the experimental methodology to either provide support for individual counterarguments against Libet's conclusions (e.g. Keller & Heckhausen, 1990; Trevena & Miller, 2002; Schurger, Sitt, & Dehaene, 2012; Verbaarschot, Farquhar, & Haselager, 2015) or expand the study using more advanced technology and procedures (for example using an fMRI machine, see Lau, Rogers, Haggard, & Passingham, 2004; Lau, Rogers, & Passingham, 2006; Soon, Brass, Heinze, & Haynes, 2008).

The more recent the studies are, the more they seem to be focused on specific aspects of the original Libet's experiment. For

* Corresponding author.
E-mail addresses: tomas.dominik01@upol.cz (T. Dominik), daniel.dostal@upol.cz (D. Dostál), martin.zielina@lfmotol.cuni.cz (M. Zielina), jan.smahaj@upol.cz (J. Šmahaj), zuzana.sedlackova@upol.cz (Z. Sedláčková), roman.prochazka@upol.cz (R. Procházka).

https://doi.org/10.1016/j.concog.2018.07.004
Received 31 March 2018; Received in revised form 21 June 2018; Accepted 6 July 2018

example, a study made by Danquah, Farrell, and O'Boyle (2008) aimed at the tactile stimulation which Libet used in his "S series" (see Libet et al., 1982, p. 325; Libet et al., 1983, p. 625) and showed that the tactile stimuli may be consciously registered with different latency than visual or auditory stimuli.

Trevena and Miller (2002) presented two experiments focused mainly on analyzing a type of event-related potentials called lateralized readiness potentials (LRPs) instead of the classical readiness potentials (RPs) used in Libet's case. However, these authors modified the original design and ultimately had to admit that the participants' introspective reports of the timing of conscious decision to move differed substantially from what was found by other researchers, including Libet, and that some of these W times were reported after the movement was initiated.

Schurger et al. (2012) introduced a modification, which they called "Libetus interruptus" design, and showed that the early readiness potentials' onsets observed by Libet and his team may be caused by spontaneus fluctuations in neural activity occasionally building up to a movement execution. Additionally, the W reports in their case were similar to what Trevena and Miller (2002) found (and thus different from Libet's original results).

Other studies focused on biases in reporting the introspective impressions in Libet-style experiments (e.g. Pockett & Miller, 2007; Banks & Isham, 2009; Pockett & Purdy, 2011; Dominik et al., 2017). Some of these showed that certain types of introspective reports (especially the reports of the urge to move) are highly susceptible to being distorted due to changes in the experimental situation.

These and other similar studies are extremely informative for isolated aspects of Libet's experiment. Nevertheless, their experimental designs are usually notably reduced and do not reflect the complexity of Libet's original study, which suggests that it might prove useful to conduct a complex replication study. That means a study which does not aim to challenge Libet's results or interpretations, but instead attempts to conduct the original experiment following Libet's methodological directions as closely as reasonable.

That is the aim of the present paper. Our procedure consisted of four general steps as follows: (1) we familiarized ourselves with Libet's original methodological papers (Libet et al., 1982, 1983; Libet, 1985), (2) we devised our own technical plan of the experiment following the original directions, (3) we enhanced the design slightly to overcome some of its original methodological limitations and to adjust it to equipment available to us, and finally (4) conducted the experiment including the data analysis. We aim to publish the research data along this paper to allow other researchers to revise an authentic data sample (see Dominik et al., submitted for publication).

One could argue that our study cannot be called a replication in the strictest sense, since we decided to make some changes to the design. While this might certainly be true, in case of such a complex experiment, it is often difficult to balance replicative accuracy and methodological generalizability, ultimately forcing the researchers to make choices between keeping the design intact, but less valid, and improving it so that it is more valid, but less accurately reproduced. In short, while we are aware that modifying the design may lead to changes in the results, we found some modifications necessary, either for technical or for methodological reasons.

Before reading on, we strongly recommend the reader who is not familiar with the details of Libet's experiment to read the original papers (Libet et al., 1982, 1983; Libet, 1985). While doing so, we should dedicate a few lines to suggest possible reasons why Libet and his team published the experiment in three separate articles. The answer may be that the authors simply wanted to separate different types of conclusions, as the experiment is notably complex (which is evident from the length of this paper). The first study, published by Libet et al. (1982), emphasized the analysis of the readiness-potentials (RPs). They showed, among other findings, what RPs look like before a spontaneous movement compared to pre-planned movement and that it does not occur before a skin stimulus. The second paper, published by Libet et al. (1983), introduced the introspective reports such as the moment of the first conscious urge to move (called "W" as in "wanting") or the subjective impression of the actual initiation of the movement (called "M" as in "movement"). The authors pointed out that the RP onsets generally precede not only the movement itself, but also the conscious awareness of wanting to move. The third paper (Libet, 1985) contains mainly discussion and further notes on the results, but it also introduces the concept of a conscious veto and suggests that some series in the original experiment were intended to require participants to deliberately veto an intended movement (Libet, 1985, p. 538).

2. Materials and methods

2.1. Participants

Originally, Libet worked with 6 participants (5 of them were females). These were all right-handed college students divided into two groups of three. Libet studied the second group a few months after the study of the first group (Libet et al., 1982, p. 323; Libet et al., 1983, p. 624).

Our research sample consisted of 8 participants. This increase in sample size is not large, but the number of 8 participants has a rational reason – it allows a complete rotation of three experimental conditions, which could not be satisfyingly rotated if we examined 6 participants only (see Section 2.5). Our participants were recruited from undergraduate psychology students during November and December 2015.

The recruitment consisted of three steps. In the first step, the potential respondents reacted to an offer distributed via university email and webpage by filling out a questionnaire containing items such as name, contact information, gender, handedness, age, approximate hair length (relevant to the quality of prospective EEG signal measurement), near-sightedness and so on. In the second step, we chose 13 participants with convenient answers in the questionnaire and invited them to a group briefing session, in which we introduced them to the research idea, time requirements, ethical regards and a financial reward. To the question whether the students can search for more information on the original experiment we replied that we prefer to introduce the participants gradually

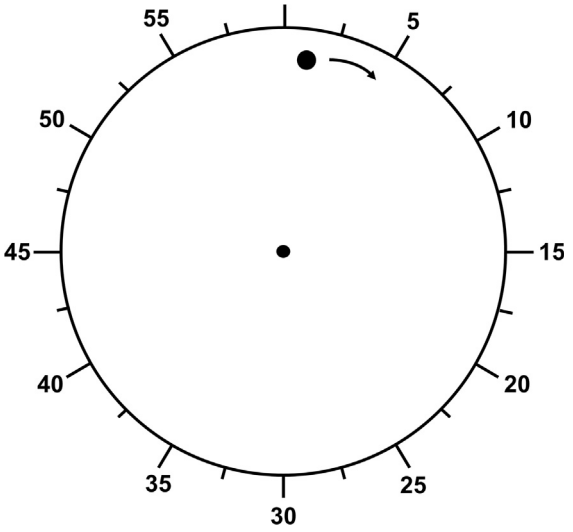

Fig. 1. User interface design. The arrow depicts the direction of the spot movement and was not actually displayed. The image presented is inverted in color, as the experiment used a white outline on black background.

ourselves. However, all participants were clearly instructed, informed and trained before any measurements were taken, so the participants' awareness of the details of the experiment should not have a major effect on the results. In the final step of the recruitment, we randomly chose 8 participants to obtain data from, with respect to our methodological requirements (4 females, 4 males, right-handed, full-time students only).

These 8 students were assigned into two groups of 4. The first group was studied from January to March 2016; the second group was studied from March to May 2016. The interval between these two study periods was two weeks.

2.2. Introspective data measurements

To obtain the introspective data, we used an office computer and a laptop, both running a custom web-based program, which we called "Libet's clock". The program is a traditional rotating-spot method designed to be used in various Libet-style experiments, but its properties strictly follow Libet's original recommendations (Libet et al., 1982, p. 324; Libet et al., 1983, p. 625). Since Libet used a CRO display, which had a bright green outline of a clock face displayed on dark background, our clock face was displayed as a white outline on black background to maximize the contrast. The outline was labeled by 12 large marks, 11 of which were numbered by increments of 5. In the middle of each pair of adjacent numbered marks, a smaller mark was displayed (see Fig. 1). A dot on the clock's circumference moved clockwise and completed one revolution in 2560 ms. The clock diameter and its distance from participant's eyes were adjusted, so that the viewing angle never exceeded 1.8° (in fact, we aimed to keep the viewing angle at exactly 1.8°).

The computer program first loaded an introduction page with the following option settings: session ID, participant's name, gender, session scenario, notes, clock diameter, clock speed (always 2560 ms per revolution) and some other options, which were, however, set always the same (such as right-handedness). Once the introduction page was filled out, an instruction was displayed (the text varied based on the session scenario and participant's gender, see the Instructions Supplement). When the participant finished reading the instructions, he or she pressed and held the mouse button, initiating the following task and marking the beginning of the task in the recordings (see Section 2.3).

The tasks varied based on the session scenario (see Section 2.5). Generally, the participants were asked to either make a movement (click the left mouse button) or wait for a skin stimulus to be delivered while watching the center of the running clock. After either of these specific events occurred, the moving dot on the clock face continued moving for a *continuation interval* (see Libet et al., 1983, p. 626) of random length ranging from 500 to 800 ms after the event, and then disappeared. Thereafter, participants were asked to report the location of the moving dot when they registered a specific introspective impression. This report could be made by two different *modes of recall*, as proposed by Libet et al. (1983, pp. 626–627). The *absolute* mode of recall (A) represents a straightforward approach – in our design, participants simply clicked on a specific spot on the clock face where the dot was when their impression first occurred. The *order* mode of recall (O) used a more complex way of reporting – after the continuation interval, the moving dot jumped to a specific *stop time*. Its position was chosen from 41 possible values[1] from the *stopping range*, which was spread across an interval from 400 ms before the event to 200 ms after the event. The participants were then asked to compare the timing of the subjective impression with the current dot's location. In this arrangement, three eventualities might occur: (1) the participants' impression comes before the stop time ("awareness first", scored as 1 point), (2) after the stop time ("clock first", scored

[1] In Libet's original experiment, only 40 measurements were taken in a single series, while one the 41 possible stop times was omitted.

as 0 points) or (3) exactly at the stop time ("together", scored as 1/2 point). Participants stated their answers by clicking on one of three boxes labeled by these three eventualities (the scores were not displayed). The final average time of the awareness was then calculated using a formula used by Libet (Church & Cobb, 1971, in Libet et al., 1983, p. 629):

(*upper end of stopping range*)–(*time interval between stop times*) × (*number of points*–1/2)

2.3. Recordings and skin stimulation

Besides measuring the introspective reports, Libet also recorded physiological data, namely EEG and EMG (Libet et al., 1982, 1983). EOG had also been employed in the first few sessions to account for unwanted eye movements, but was later discontinued, because the EOG potentials rarely occurred, as it seemed that participants could satisfyingly fixate a mark in the center of the clock (Libet et al., 1982, p. 323). Therefore, we decided to record EEG and EMG only, as the EOG would presumably pose unnecessary additional discomfort to our participants.

For both the EEG and EMG recordings, we used the BIOPAC MP150 unit with EEG100C and EMG100C amplifiers. The EEG recordings were taken to obtain data on the readiness-potential onsets, which are one of the central elements of Libet's findings (Libet et al., 1982, 1983; Libet, 1985). We recorded the EEG from six standardized 10–20 locations using 0.1 Hz high-pass and 35 Hz low-pass filters. The Fp_1, Fp_2, C_z and P_3 electrodes were recorded in accordance with Libet's directions (Libet et al., 1982). However, Libet also recorded additional non-standard electrodes C_c and C_i (Libet et al., 1982, p. 323), which are not included in the 10–20 system. We decided to record C_3 and C_4 electrodes instead, because the EEG cap available to us does not allow recording of un-standardized locations outside of 10–20 system. The electrodes were embedded in the BIOPAC CAP100C electroencephalography cap filled with an electroconductive gel. Reference electrodes were placed on both left and right ear lobes.

Because the movement, which our participants were asked to conduct, was a mouse click, we recorded the EMG from the musculus extensor indicis using two EL503 Ag/AgCl electrodes applied to the skin abraded with ELPREP gel. The ground electrode was located on the upper part of the musculus brachioradialis. We applied 10 Hz high-pass and 500 Hz low-pass filters.

Both the EEG and EMG data were obtained, stored and pre-processed in the AcqKnowledge software v4.4. The AcqKnowledge also recorded the mouse clicks as a digital input reaching 1 when the mouse button was being pressed and returning to 0 when it was released. This was achieved by connecting the computer running Libet's clock to the BIOPAC STP100C module via a CBL110C cable plugged into the LPT port of the computer. To convert a click to an output signal in the LPT port, we used a simple custom script. This arrangement was utilized to accurately align the timeline in Libet's clock with the timeline of the physiological recordings, as each click also created a timestamp in Libet's clock.

Another important part of Libet's original experiment was the skin stimulation (S) series (Libet et al., 1983, pp. 627–628) employing an electrical stimulator placed on the back of participant's hand (Libet et al., 1982, pp. 323–324). We used a non-electrical tactile stimulator TSD190 provided by BIOPAC connected to the STM100C module (set on CH15 source setting). The stimulator was placed on the anterior side of participant's left wrist. It was controlled by an analog signal sent from a laptop via a 3.5 mm audio cable to channel 16 on the UIM100C module. For safety reasons, we used a battery-powered laptop (connected to the same display which was used in all other types of tasks) instead of an office PC. The stimulus intensity was adjusted at the beginning of each session by reducing it to a level at which participant could no longer feel the stimulus and subsequently increasing the intensity slightly (by about 15% of a turn of the cylindrical slider on the STM100C). The stimulus was thus "*sufficiently weak to make recognition somewhat difficult, but sufficiently above threshold to eliminate equivocation about stimulus delivery*" (Libet et al., 1982, p. 324). The temporal alignment of Libet's clock with the AcqKnowledge timeline was in this case achieved using the skin stimuli themselves, because the STP100C module, otherwise registering the mouse clicks, had to be disconnected from the rest of the system for technical reasons.

2.4. Types of the experimental tasks

In Libet's original experiment, each participant attended 6–8 half-day *sessions*. Intervals between two following sessions were about one week long. The first, and, in some cases, the second, session was dedicated to training (*training sessions*); others were called *regular sessions*. Each of Libet's sessions consisted of a certain number of *series* of various types, which consisted of 40 individual *trials* each. Each series was also preceded by 10 retraining trials (Libet et al., 1982, p. 325).

Libet classified the series into categories, but used different classifications in Libet et al. (1982) and Libet et al. (1983). Moreover, new information was brought in Libet (1985). In this section, we will unify the terminology and create a system of series labels, which will be used in the description of our procedure. In Libet et al. (1982), the series was divided into three types: (1) *Self-initiated voluntary acts*, (2) *Pre-set motor acts* and (3) *Skin stimuli at unknown times*.

The *Self-initiated voluntary acts* series corresponds to the W and M series introduced in Libet et al. (1983), but the W and M tasks are not identical. Therefore, we will use the **W** and **M** labels to describe these two slightly different procedures.

Pre-set motor acts are introduced in Libet et al. (1982), but not mentioned in Libet et al. (1983). Furthermore, a new type of pre-set series "with conscious veto" is introduced in Libet (1985, p. 538). Therefore, we will label the pre-set series without vetoing the movement as **P** and the pre-set series with vetoing the movement as **Pv**.

Skin stimuli at unknown times series – also called S series in Libet et al. (1983) – is the only series with clear meaning throughout the original papers, and will be labeled **S** in this text.

In the next segment, we will introduce a brief description of each of the series types and explain our methods of their execution:

The **M series** in our replication consisted of 40 trials in mode A or 41 trials in mode O. During each trial, the participant was

seated in a medical armchair. A wooden board was placed on the chair's armrests and a computer mouse was put on top of it. The participant laid his or her right hand on the mouse and fixated the mark in the center of Libet's clock. The participant was asked in the textual instruction to click the left mouse button whenever he or she felt like it (see the Instructions Supplement). The click was followed by the continuation interval after which a prompt was displayed, asking the participant to report the time of subjective recall of the beginning of the movement (M). The prompt was different for A and O modes of recall (see Section 2.2). Participants were advised to look at a wall after making the report, if they felt the need to relieve their eyes; and then click an OK button to continue. Participants were also instructed not to blink while the clock was running, unless the urge to blink became uncomfortable – in such a case, participants could blink, but then had to wait at least one whole clock revolution before making another click.

The **W series** was almost identical to the M series, with a slight difference in the introspective event reported after the click. In the M series, participants reported when they first realized that they were moving their finger. In the W series, participants reported when they realized the first urge to move (W). All other instructions were kept the same.

In Libet's original experiment, the movement made by the participants was a flexion of their wrist or fingers (Libet et al., 1983, p. 625). We decided to use a mouse click, mainly because our technical setup did not allow us to control the computer using EMG activation; hence it was impossible to stop the Libet's clock if no other input than EMG was provided. However, we assumed that this trade-off might bring some advantages. First, as a lot of recent Libet-style experiments employed a mouse click or a keypress as a substitute for the movement onset measured with EMG (e.g. Pockett & Miller, 2007; Soon et al., 2008; Verbaarschot et al., 2015; Caspar & Cleeremans, 2016; Dominik et al., 2017), we aimed to compare the time of a mouse click to an EMG onset. This was already done by Haggard and Eimer (1999) who found that the keypress typically occurs 30–50 ms after the EMG onset, and we aim to verify their result. Second, the mouse click should be a far more abrupt and better bounded movement than a flexion of a wrist or fingers – it should therefore be easier for the participants to determine the timing of their movement. Nevertheless, clicking a mouse button brings another problem that needs to be resolved. Banks and Isham (2009) showed that a delayed auditory feedback to the movement might systematically distort the W reports. Because the mouse click actually provides this auditory feedback (presumably delayed by 30–50 ms compared to the EMG onset, as Haggard and Eimer, 1999, suggest), it seems advisable to eliminate this factor. We managed that by putting soft earplugs into participants ears (after making sure that the participants do not have any question to be answered, of course). All participants reported that they were unable to hear the click and that they did not feel too uncomfortable while having the earplugs in their ears.

The **S series** also consisted of 40 or 41 trials for A and O modes of recall, respectively. Libet's reason to include this type of series was to have a "correction" for how accurate participants' time perception was (Libet et al., 1983, p. 627). In our case, participants were seated in the same position as in the W and M series, but as the event was a skin stimulus and not a mouse click, EMG was not recorded. Instead, a tactile stimulator was placed on the anterior side of participant's left wrist (see Section 2.3), which was resting on the wooden board in front of the participant in supine position. The participant was instructed to sit calmly, fixate the mark in the center of the clock and wait for the stimulus. The skin stimuli in the original experiment were delivered by an experimenter at random times unknown to the participant, but never during the first revolution of the clock (Libet et al., 1982, p. 325; Libet et al., 1983, p. 625). We used an algorithm running within our "Libet's clock", which sent an analog signal driving the stimulator at random times after the first revolution was completed. After the continuation interval following the stimulus, a response prompt was displayed asking about the time in which the participant registered the stimulus (again, the prompt and the way of responding varied for the A and O modes of recall). As in the W and M series, participants were advised to relieve their eyes if needed, by looking at a wall before continuing the series.

The **P series** was similar to the W and M series in terms of experimental setting, but the task was different. Participants were sitting with their right hand on the mouse (EMG was recorded) and watched the center of the clock. In this case, there was a fixed bright green dot on the clock's circumference in addition to the moving dot. This green dot (the target point) appeared on a random ("pre-set") position, which was different in each trial (as opposed to Libet's original design in which the target mark was placed on the same spot in each block of 10 successive trials). The participant's task was to click the mouse button when the moving dot reached the green mark's position, with maximum accuracy possible. After the click was made, no response prompt was displayed; instead, the target point disappeared, while the white dot continued moving for another few revolutions, after which a new target point was displayed at another location, so the task could be repeated. If a participant missed the target point, he or she was instructed to simply wait for the next revolution and then try again; thus, the clock did not force any pace on the participants. There were always 40 trials in the P series.

The **Pv series** was almost identical to the P series. The difference was that the participants were instructed to prepare to make the movement (click the mouse button) exactly at the time marked by the target point, but then stop ("veto") the movement just before it begun (for more details, see the Instructions Supplement). If the participant believed that the task was done right, he or she clicked the mouse button at any time during the next revolution to give a signal that the veto was made. The target point then disappeared and reappeared in another location a few seconds later. If the participant was unsure that the veto was properly performed, he or she simply waited for another revolution without clicking and then tried again. As in the P series, there were always 40 trials in each Pv series.

The participants were initially trained in all these types of sessions. Besides the initial training sessions (see Section 2.5), each series was preceded by 10 training trials consisting of the same task and employing the same mode of recall as the following 40 or 41 trials. This was done with respect to Libet's directions (Libet et al., 1982, p. 325) to re-familiarize the participants with the upcoming task. In Libet's original design, there was also an S series of 25 trials placed at the beginning of each regular session, which were also intended to train the participants – in this case to give accurate introspective reports. Libet did this by providing feedback on how close their responses were to the times of actual stimuli every 5 trials (Libet et al., 1983, p. 628). Additionally, Libet provided the

Table 1
An overview of one of the experimental arrangements. The W, M, S, P and Pv indicate the series type, (A) and (O) indicate the absolute and order mode of recall, respectively. The gray fields mark the training session or the re-training 10 trials series.

Session 1	*Personal and medical history. EMG biofeedback. Spontaneous movement training*					
training	*Explanation and training: M(A), W(O), S(A), S(O) series; 10 trials each*					
Session 2	re-training	W(A)	re-training	M(A)	re-training	S(A)
regular	W(A)	(40 trials)	M(A)	(40 trials)	S(A)	(40 trials)
	(10 trials)		(10 trials)		(10 trials)	
Session 3	re-training W(O)	W(O)	re-training M(O)	M(O)	re-training	S(O)
regular	(10 trials)	(41 trials)	(10 trials)	(41 trials)	S(O)	(41 trials)
					(10 trials)	
Session 4	re-training M(A)	M(A)	re-training W(A)	W(A)	re-training	S(A)
regular	(10 trials)	(40 trials)	(10 trials)	(40 trials)	S(A)	(40 trials)
					(10 trials)	
Session 5	re-training M(O)	M(O)	re-training W(O)	W(O)	re-training	S(O)
regular	(10 trials)	(41 trials)	(10 trials)	(41 trials)	S(O)	(41 trials)
					(10 trials)	
Session 6	re-training	P	re-training	Pv	re-training W(A)	W(A)
supplementary	P	(40 trials)	Pv	(40 trials)	(10 trials)	(40 trials)
	(10 trials)		(10 trials)			
Session 7	re-training	Pv	re-training	P	re-training W(O)	W(O)
supplementary	Pv	(40 trials)	P	(40 trials)	(10 trials)	(41 trials)
	(10 trials)		(10 trials)			

feedback at the end of each regular S series (Libet et al., 1983, p. 627). Libet believed that this introspection accuracy training was crucial, because a systematic bias in the S series would also apply to other tasks, such as the W series (he also suggests a correction for the W and M reports by subtracting the S time, see Libet et al., 1983, pp. 630–631). However, Gomes (1998, p. 590) points out that this feedback training performed on a regular basis throughout the experiment might lead to variable results across the sessions. Additionally, there seems to be no reason to use the training based on the skin stimulation specifically, because Danquah et al. (2008) showed that the introspective reports differ for tactile, visual and auditive modality – this implies that the reports provided after an auditory feedback training would presumably differ from those provided after a tactile feedback training. With respect to these objections, we chose to withdraw from following Libet's directions in this case, and decided to refrain from providing any feedback in the S series and to remove the 25 initial S trials from all sessions in our design. While being aware that this step might alter our results significantly compared to Libet's original outcomes, we believe, that in this case it is more useful to conduct the experiment in a way, which is arguably more valid than replicatively accurate.

2.5. Progression of the experiment

As stated above, we conducted all the task variants, which were present in Libet's original design. We organized the series into 7 sessions. To emphasize the difference between the two distinctively different types of sessions, we introduced a new label *Supplementary session* for the sessions with the P and Pv series. In this section, we will introduce the content of each session and explain how the experimental conditions were rotated. See Table 1 for an overview of one possible experimental progress.

The first session was a **training session** and included taking personal and medical history, an EMG biofeedback, spontaneous movement training and familiarization with the W, M and S tasks including learning the difference between the A and O modes of recall. There was only one training session for each participant, as we found no reason to repeat the training for the second time.

The personal and medical history inquiry focused on potential contraindications of EEG and EMG recording or sources of EEG artifacts, such as epilepsy, psychopathology, head injuries or relevant medication. None of the participants reported any of the potential risks.

The purpose of the EMG biofeedback was to train the participants to click the mouse button abruptly enough so that the EMG activation reaches its maximum in the shortest time possible. This was done using the BIOPAC MP36 unit and the lesson template L01 – Electromyography (EMG) I in the BIOPAC STUDENTS LAB software, which contains two tasks suitable for our needs. The first of these tasks required recording and visualizing the EMG in a graph – this helped participants learn in real time what the recording looks like and how to click the mouse button so that the EMG activation is fast enough. The second task was to listen to headphones, in which the EMG signal was transformed into a form of a sound wave – this was intended to further help participants to make abrupt clicks, so that the rise in the sound volume was as short as possible. If the experimenters present to the following sessions noticed that the EMG rise tended to be insufficiently steep, they notified the participant and asked him or her to make the click more abruptly.

After it was clear that the participant knew how to click "correctly" (which the participants typically achieved in 30–40 min), we presented Libet's clock to the participants. The experimenters explained its function and then presented four examples of tasks relevant to the following 4 regular sessions. The first presented task was M(A) – M series in absolute mode of recall – which we found the simplest to explain, followed by the W(O), S(A) and S(O) series. The participants were assured that the A and O modes of recall will not vary during a single session and that the changes in the present session are intended for instructional reasons only. During the M(A) and W(O) series, the participants were instructed to make the movement spontaneously, with no pre-planning, just as the urge appeared (this was in fact stated in all relevant textual instructions in the following sessions, see the Instructions Supplement).

Also, the technical equipment was shown, described and demonstratively used during the training session, so that all participants were familiar with as many aspects of the experiment as possible. Special care was given to whether the participants feel comfortable with the equipment (i.e. the EEG cap, the EMG electrodes, the skin stimulator or the earplugs). However, no explicit statement about Libet's results or conclusions was made until the end of the last session.

The **regular sessions** consisted of three 10 trials re-training series and three regular series in the W, M and S conditions. Before each of the series, a separate textual instruction was displayed and the participants were encouraged to read it every time to remind them of the current task and some principles, which were constant during the experiment (such as the blinking rules, fixation mark, spontaneity of the movement etc.). In the first several sessions, the participants were required to repeat the instruction in their own words, so that it was clear that they understand.[2] After verifying that the participant comprehends the task, the first series (either W or M) begun. After the first 100 or 102 trials (10 re-training W, 40 or 41 regular W, 10 re-training M, 40 or 41 regular M) were completed, two questions were asked:

1. "Did you notice during the experiment that you felt that the movement was not spontaneous, that you pre-planned it?"
2. "Did it occur to you during the experiment that you were surprised by the movement? That it came on its own, without you knowing?"

If the participant answered positively to any of these questions, he or she was asked in which series it was so and the answer was recorded in the protocol (see below).

Once the W and M series were finished, a change in the technical setting was made to prepare for the S series (see Section 2.3). Then, another 50 or 51 trials with skin stimulation were conducted (10 re-training S, 40 or 41 regular S).

The sixth and seventh sessions were the **supplementary sessions** and differed from the regular ones by introducing the P and Pv series. At the beginning of the sixth session, the participants were given training in the P and Pv tasks, which was similar to the training of the W, M and S series in the initial training session. Once the participants were comfortable with the new task, they were presented with three series – P, Pv and a supplementary W series, which did not differ from the corresponding series in the previous sessions (its inclusion was intended to increase the number of the W trials, as these seem to be the most crucial aspect of Libet's experiment). This scenario was repeated in the seventh session, but with reversed order of the P and Pv series and opposite mode of recall for the W series.

The experimental conditions in the regular and supplementary sessions were rotated based on several principles:

1. The mode of recall stayed constant during each session.
2. The mode of recall alternated in consecutive sessions (both regular and supplementary).
3. The order of the M and W series was reversed between the 3rd and the 4th session.[3]
4. The order of the P and Pv series was reversed between the 6th and the 7th session.
5. The order of the P and Pv series was independent of the mode of recall in the following W series.
6. Every possible combination was assigned to two participants (one on the first group and one in the second group).

These principles resulted in a schedule depicted in Table 2.

Since two experimenters out of five attended every session, a standardization of the procedure was needed. Therefore, we employed textual **protocols** – documents containing step-by-step directions on how to conduct each session, describing the most specific details, such as which connectors should be placed into which socket or which information should be entered in Libet's clock settings. The protocols were participant- and session-specific, which means that no pair of the 56 protocols was identical. Each protocol also contained several blank fields for procedural and technical notes and participants' answers to the standardized questions (see above).

2.6. Data analyses

The data were acquired from multiple sources in different forms and needed to be unified in a single framework. We aimed to use the R software for the analyses and therefore needed all the data converted into compatible formats.

The introspective and technical data from Libet's clock (i.e. the subjective reports of the M, W and S times, the timing of the skin stimuli in the S series and the data on when a click occurred in the M, W, P and Pv series) were exported as CSV files from the database, to which the Libet's clock software had sent all the data. The physiological recordings (one EMG and six EEG channels) and other data obtained using the MP150 unit (the keypress channel and the stimuli channel) were exported as CSV files from the AcqKnowledge software. These two types of CSV files were merged and temporally aligned based on the mouse clicks or the skin stimuli, which were recorded by both Libet's clock and the MP150 unit (see Section 2.3). There were two merged files for each regular session (one for the M and W series, one for the S series) and one merged file for each supplementary session (including all the P, Pv and W series). The merged files were then segmented into epochs spreading from 2500 ms before the event (mouse click or skin stimulus) to 800 ms after the event.

[2] Once the experimenters were sure that a participant has a grasp of the various variants of the procedure, the repetition of the instructions was no longer required.

[3] This was not fully achieved in one case due to experimenters' error. See the asterisk in Table 2.

Table 2
An overview of the rotation of experimental conditions. The participants 1, 2, 3 and 4 were members of the first group studied in the first two months; the participants 5, 6, 7 and 8 were studied in the following two months. The W, M, S, P and Pv indicate the series type, (A) and (O) indicate the absolute and order mode of recall, respectively. The asterisk indicates that in the case of the 7th participant an error occurred and the order of the sessions was altered (namely, the 3rd and the 5th sessions were swapped).

	Participant 1 and 5	Participant 2 and 6	Participant 3 and 7*	Participant 4 and 8
Session 1 *training*	EMG biof. – M(O) – W(A) – S(A) – S(O)	EMG biof. – M(O) – W(A) – S(A) – S(O)	EMG biof. – M(O) – W(A) – S(A) – S(O)	EMG biof. – M(O) – W(A) – S(A) – S(O)
Session 2 *regular*	W(A) – M(A) – S(A)	W(O) – M(O) – S(O)	M(A) – W(A) – S(A)	M(O) – W(O) – S(O)
Session 3 *regular*	W(O) – M(O) – S(O)	W(A) – M(A) – S(A)	M(O) – W(O) – S(O)	M(A) – W(A) – S(A)
Session 4 *regular*	M(A) – W(A) – S(A)	M(O) – W(O) – S(O)	W(A) – M(A) – S(A)	W(O) – M(O) – S(O)
Session 5 *regular*	M(O) – W(O) – S(O)	M(A) – W(A) – S(A)	W(O) – M(O) – S(O)	W(A) – M(A) – S(A)
Session 6 *supplementary*	Pv – P – W(A)	Pv – P – W(O)	P – Pv – W(A)	P – Pv – W(O)
Session 7 *supplementary*	P – Pv – W(O)	P – Pv – W(A)	Pv – P – W(O)	Pv – P – W(A)

The EMG data were analyzed to find the onset of the EMG activation (which we will call the **EMG_0**), which serves as a reference point for temporal analyses of the W, M and P series. The EMG data was processed in three steps. In the first step, we calculated the absolute value of each EMG waveform. In the second step, the Butterworth procedure was applied to rectify the data. The third step was to find the EMG onset, which we did by calculating a threshold value equal to five times the IQR (interquartile range) of a segment from −1500 to −300 ms added to the respective Q3 (upper quartile); this threshold was subsequently applied to the segment of −300 to +200 ms (we stretched this range into the positive values, because in some cases in the P series the participants were expected to exceed the zero time, which in this case was marked by the position of the target point).

The EEG data were treated separately for each channel. Before any further analyses, we rejected those EEG recordings, which contained excessive noise (284 recordings in 68 series across all channels, sessions and participants, out of 858 recordings in total). The remaining recordings were additionally filtered (using 0.5 Hz high-pass and 35 Hz low-pass filter) to reduce or eliminate long polarizing shifts of the EEG baseline and the residues of the 50 Hz noise. These filtered recordings were segmented into 34,656 epochs where every epoch represented one trial and one electrode. Epochs from the W, M and P series containing no EMG onset or an invalid EMG onset and epochs from Pv series containing an EMG onset were rejected (5640). Epochs, which contained blinking artifacts and artifacts caused by eye movement were further discarded if the voltage P-P range exceeded a rejection value of 0.15 mV (11,347).

Every remaining epoch (17,669) was then temporally locked to the respective **reference point**: the EMG onset in the W and M trials, the skin stimulus in the S trials and the target point position in the P and Pv series.[4] Each channel in each series was then averaged. Averages calculated from less than 10 trials (22) were further rejected. The averaged data line plots were used to represent the mean course of activation on a given electrode before and after the reference point. Before further analyses, grand averages were calculated (i.e. we plotted the mean courses of activation based on all 17,669 valid trials in respective series regardless of the session, but with respect to the electrode). To assess the RP onsets in individual series, we used both methods suggested by Libet et al. (1983, p. 632): (1) the MN (main negative) method and (2) the $RP_{90\%}$ method.

The **MN method** consists of an eye-ball inspection of each graph by multiple independent investigators. For our analyses, we employed five members of our team, who were all aware, what the onset should look like (we designed a custom "Guide to the RP identification", see the corresponding supplement), but who were not aware of the type of series a respective graph was obtained in. An exception was one of the five researchers who also merged the estimates into one file and who was aware of the respective series types at the time. The conclusions were then confirmed by one independent examiner who was not a member of the authors' team. While estimating whether an RP is present, we proceeded rather conservatively (i.e. we did not identify a negativity as an RP if the waveform lacked some important characteristics, such as adequate artifact-free straight baseline).

The **$RP_{90\%}$ method** is a more objective approach to assessing the RP onset. It uses a calculation of the area under the RP curve preceding the reference point (see Libet et al., 1983, pp. 632–633). Unlike Libet, we did not have to take geometric measurements from a paper tape, as we used a computer algorithm based on Libet's calculations instead. The algorithm was applied to the averaged EEG plots. A baseline was calculated for each plot as an average of the EEG waveform in the interval −1500 to −1000 ms in the W, M and S series and −2500 to −1500 ms in the P and Pv series.[5] A window 50 ms wide started with its right edge aligned with the

[4] This is because in the Pv series, an EMG onset is not supposed to be found (as the participant is told to veto the movement). Therefore, we analyze the EEG and calculate the RP onsets in the Pv series relative to the target point. We performed the same procedure with the P series to make the RP onset times in the P series comparable to those in the Pv series.

[5] Later baseline (-1500 to -1000) was preferred, because gradual increases and decreases in voltage early in the baseline might shift the RP onsets; however, in the P and Pv series, the RP onsets are expected to have earlier onsets than those in the W and M series – therefore, the baseline was moved to -2500 to -1500 in these series.

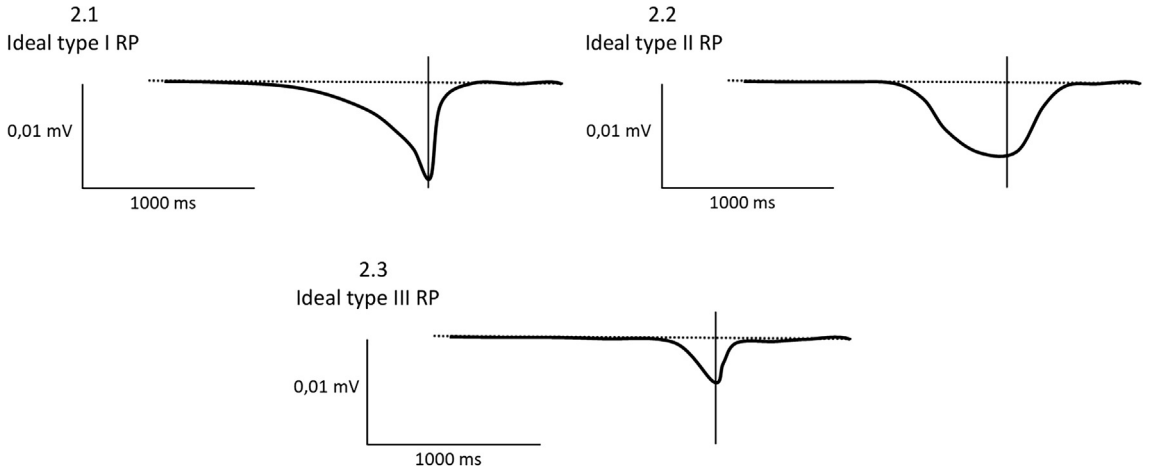

Fig. 2. Our depictions of ideal type I, type II and type III RPs. These figures do not represent any data, they are drawn by hand to illustrate the most characteristic features of the respective RP types according to Libet et al. (1982). The horizontal dotted line represents the baseline, the solid vertical line represents the EMG onset.

reference point and moved to the left (i.e. to the negative values) by steps of 1 ms. In each step, the area under the EEG waveform (or rather above it, as we plotted negative potentials bellow the baseline) was calculated. If the window entailed the curve both under and above baseline, the upper portion of the area was subtracted from its lower portion. Once the area under the curve was smaller than 12,5 μV/50 ms, all remaining data points preceding the current window position including the segment currently covered by the window were considered 0. The RP onset was placed to the lower limit of 90% portion of the area (calculated from positive to negative, starting from the reference point). Our algorithm was in fact more precise than Libet's original calculation, as it used a window of constant size.

Three researchers also assessed the RP types (I, II or III). Libet et al. (1982, p. 326) describe **type I RP** as follows: *"In type I RP a gradually or steadily rising, ramp-like form begins distinctly prior to −700 ms"*; the authors also add that more extreme examples of type I RP occurred in early sessions before the spontaneity requirement was added into the instruction. Because our instruction contained the spontaneity mention in all sessions, we expect type I RPs to be rather rare and if it occurs, we expect it to begin later. The ideal depiction of type I RP is shown in our own Fig. 2.1.

Type II RP is described as follows: "*In type II RPs, the main rise of negativity starts in the range of about −400 to −700 msec, (…) The main portion of this RP is often somewhat dome-shaped rather than ramp-like in form.*" (Libet et al., 1982, p. 326). The authors admit that the RP II may be preceded by some irregular negativity, but suggest that this negativity does not have the characteristics of the early rise of type I RP. The ideal type II RP is depicted in Fig. 2.2.

Type III RP seems to be less frequent potential charaterized as follows: "*In type III RPs, the main rise of negativity does not appear until about −250 to −200 msec (…). Total durations of any detectable negativity and especially total areas of RP are also low.*" (Libet et al., 1982, p. 326). For the image of an ideal type III RP see Fig. 2.3. In our analyses we extended the type III RP definition to any late-onset negativities that cannot be regarded type II RPs (some of our type III RPs' onsets suggested type II RP, but their shape was rather sharp, not dome-shaped – these were thus identified as type III RP).

Libet also suggested that the pre-set series elicit a special ramp-like RP with an onset earlier than −1400 ms (which was the edge of Libet's epoch range; see Libet et al., 1982, p. 330). We adopted this view, but had to extend the pre-set RP definition, because RPs in our pre-set series exhibited rather dome-shaped form (see Section 3.3).

Additionally, four of the examiners also assessed whether a P300 waveform is present, as it is expected to follow the stimulus delivery in the S series (see Libet et al., 1982, pp. 330–331; for the P300 characteristics, see e.g. Picton, 1992). The actual occurrence was then judged by the researcher who merged the individual assessment to one conclusion.

The W and M data obtained using Libet's clock in the A mode of recall were analyzed for each trial by subtracting the EMG_0 timestamp from the timestamp of participant's response. For example, if a participant made a W report at a location which corresponded to 204 ms before the EMG_0, that single W time would be −204 ms. The S reports made in the A mode of recall were analyzed similarly – the timestamp of the skin stimulus was subtracted from the timestamp of the participant's response.

The responses acquired in the O mode of recall could not be analyzed for each trial and were calculated for the whole series using the formula presented at the end of Section 2.2. The W and M reports were then corrected for the mean EMG_0 calculated from the whole respective series.

In both cases, the W reports could also be corrected by mean S report, but as explained in Section 2.4, we decided to refrain from such procedure. However, as we publish the introspective reports data (see Dominik et al., submitted for publication), we invite interested researchers to explore this idea on our dataset.

Other analyses presented in the Section 3 either use standard statistical procedures briefly described along the relevant results.

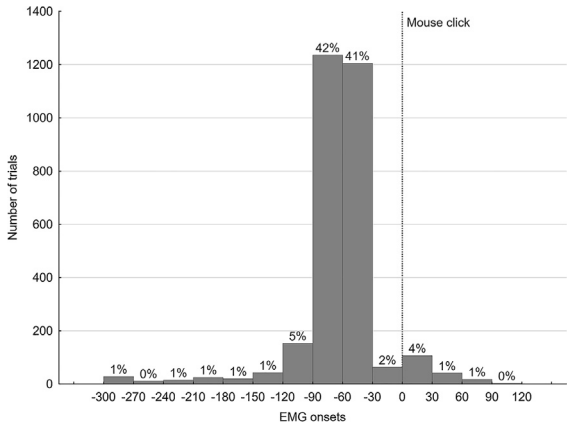

Fig. 3. The histogram of EMG onsets in the W and M series compared to the mouse click (the x-axis values signify the intervals' upper boundaries; trials with no EMG onset found discarded; n = 2968).

3. Results

3.1. EMG onset timing

The EMG onset is relevant for the W, M and P series. In the W and M series, it constitutes the reference point for the EEG onset and the introspective reports. In the P series, we are interested in the comparison of the target point, the mouse click and the EMG onset.

3.1.1. EMG onsets in M and W trials

The W and M series serve specific interpretational purpose, and thus we treated the EMG onset in these series separately from the P series. There were 3239 individual trials in 80 W or M series regardless of the mode of recall. In 271 trials, no EMG onset was found. Another 98 trials contained EMG onsets preceding the mouse click by more than 150 ms. Additionally, 169 EMG onsets occurred after the mouse click by 1 to 106 ms. We discarded all 271 trials with no EMG onset found, 98 trials with EMG earlier than 150 ms before the click and 169 trials, in which the EMG onset followed the click. We were thus left with 2701 valid EMG onsets in the range from −150 ms to 0 ms (compared to the mouse click), which were included in the RP analyses and the analyses of introspective reports.

The mean difference between the EMG onset and the mouse click in the M and W series with no trials discarded is −62.5 ms (SD = 42.8). After discarding the EMG onsets earlier than −150 ms and later than 0 ms, the mean value practically did not change (M = −62.4 ms; SD = 19.2). Haggard and Eimer (1999) assert that the typical EMG onset occurs 30–50 ms before the mouse click. In our case, only 567 of the 2701 valid results were within this range. Instead, the typical majority (2441 trials, 90% of the valid trials in range from −150 to 0 ms, 83% of all trials) occurred between −90 (excluding) and −30 ms (including). For more synoptic overview, see Fig. 3. Using the Wilcoxon signed-rank test[6] we found that in the case of the M and W trials the EMG onsets differ significantly from the mouse click [Z = 45.459, p < .001], but we also claim that the EMG onset may commonly precede the press of a button by more than Haggard and Eimer (1999) suggested.

3.1.2. EMG onsets in P trials

The EMG onsets in the P series (together with the mouse click times) show how accurate the participants were while attempting to click at the time marked by the static target point on Libet's clock. We are interested in three kinds of time differences: (1) between the EMG onset and the mouse click (as in the case of W and M series), (2) between the EMG onset and the target point and (3) between the mouse click and the target point. There were 640 trials in the P series (we do not include the Pv series). In 15 trials, no EMG onset was found. Sixteen EMG onsets occurred more than 150 ms before the mouse click, 35 EMG onsets occurred after the click. We did not discard any trials from the RP analyses, except for those with no EMG onset found, as we relate the RP onsets to the target points and not the movement onset. The mean of the EMG onsets relative to the mouse clicks including all trials was −64.6 ms (SD = 41.9); if we include only the EMG onsets later than −150 ms, but earlier than 0 ms, the mean value is −66.6 ms (SD = 21.1). For an overview of the difference between the EMG onset and the mouse click in the P series, see Fig. 4.1.

As for the participants' accuracy in the P series measured by the EMG, the mean EMG onset occurred 35.4 ms before the target point (SD = 64.4, see Fig. 4.2). If we measure the accuracy by the mouse clicks, the mean mouse click occurred 28.4 ms after the target point (SD = 67.4, see Fig. 4.3). Overall, we can state that the participants are generally accurate in clicking at the pre-set target time, with the EMG onset preceding the target time and the actual mouse click being slightly delayed.

[6] We ignore the assumption of complete trials independence in this case to keep the text simple. However, we can safely assume that the difference would remain significant even if we used much more complex statistical procedures.

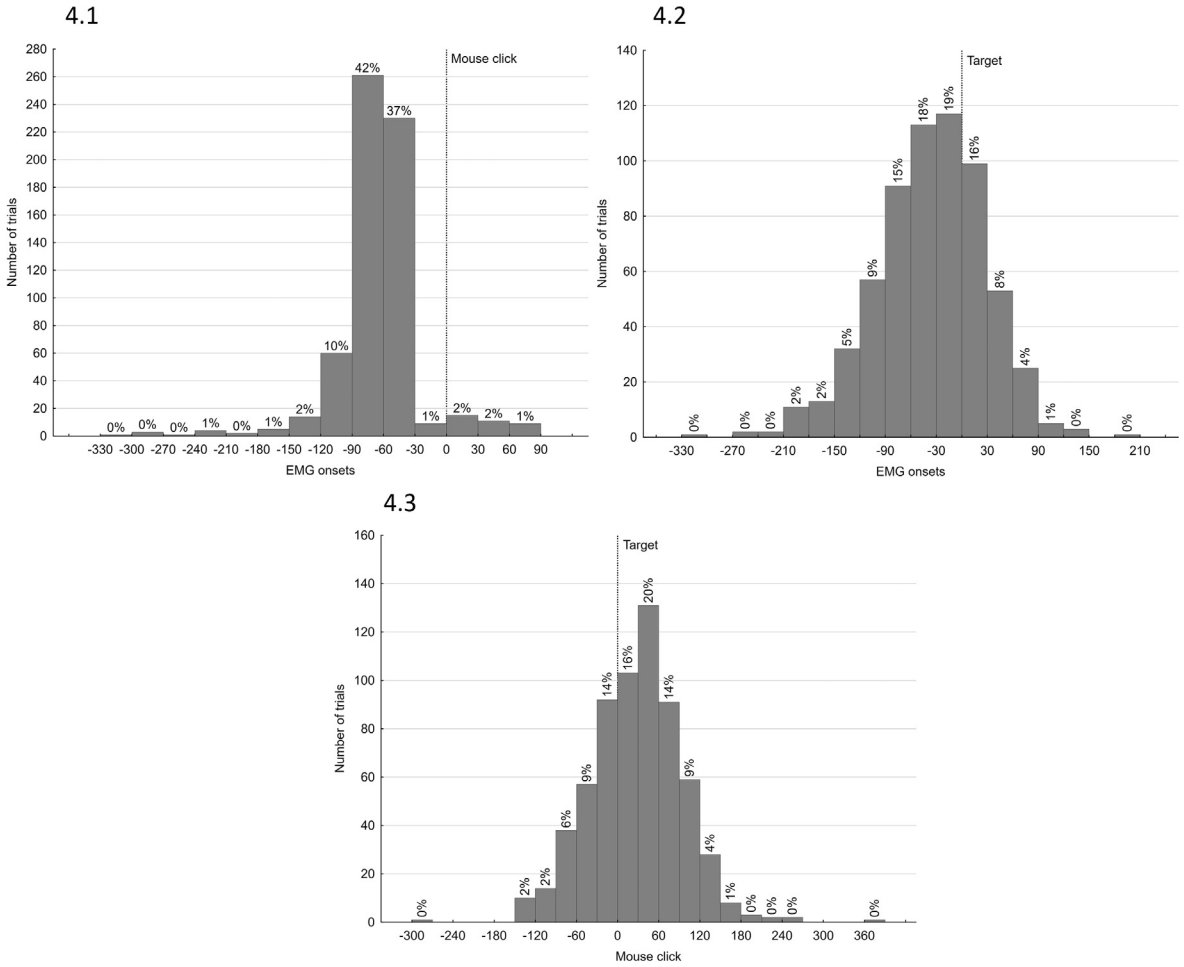

Fig. 4. The histograms of EMG onsets in the P series compared to the mouse click (n = 625, trials with no EMG onset found discarded, Fig. 3.1), EMG onsets in the P series compared to the target point (n = 625, trials with no EMG onset found discarded, Fig. 3.2) and the mouse clicks in the P series compared to the target point (n = 640, no trials discarded, Fig. 3.3). The x-axis values signify the intervals' upper boundaries.

3.2. Introspective impressions timing

All introspective reports (W, M and S) were treated separately for the absolute (A) and order (O) mode of recall (see Section 2.2). All temporal data in this section are related to their respective EMG onset or stimulus delivery (in the O mode of recall, the mean value is related to the mean EMG onset or the stimulus delivery calculated for the whole series).

3.2.1. Absolute introspective reports

The absolute reports of the spontaneous movement initiation, i.e. the M(A) reports, were collected in fifteen 40-trials series (originally, 16 series were carried out, but one M(A) series with participant 4 was corrupted due to a technical error). Only trials with a valid EMG onset in the range from −150 to 0 ms (531 out of the 600 trials) were included in the analyses. The grand averaged M time was reported slightly after the EMG onset with substantial amount of variability (M = 26.7 ms, SD = 134.7, see Fig. 5.1). The M report mean estimate based on the whole series (i.e. average of the mean reports) is M = 30.9 ms (SEM = 79.8).

The absolute reports of the first urge or wanting to move, i.e. the W(A) reports, were collected in twenty-four series with 808 out of the 960 trials containing valid EMG onset. The overall mean W time was reported about 100 ms before the EMG onset with even larger variability than in the case of the M reports (M = −98.5 ms, SD = 197.6, see Fig. 5.2); calculating using the series means – M = 101.2 ms (SEM = 151.1).

The absolute reports of the skin stimulus registration, i.e. the S(A) reports, were collected in sixteen 40-trials series with no trials discarded (the EMG recordings would be irrelevant in this case and the skin stimulator did not exhibit any significant unsystematic error, therefore no trials had to be rejected). The analysis of the 640 S(A) trials showed that the skin stimulus was reported to be registered on average about 150 ms after the stimulus (M = 146.0 ms, SD = 150.2, see Fig. 5.3); calculating using the series means – M = 146.0 ms (SEM = 77.3).

Fig. 6 contains an overview of the means and confidence intervals of the M(A), W(A) and S(A) reports for individual participants.

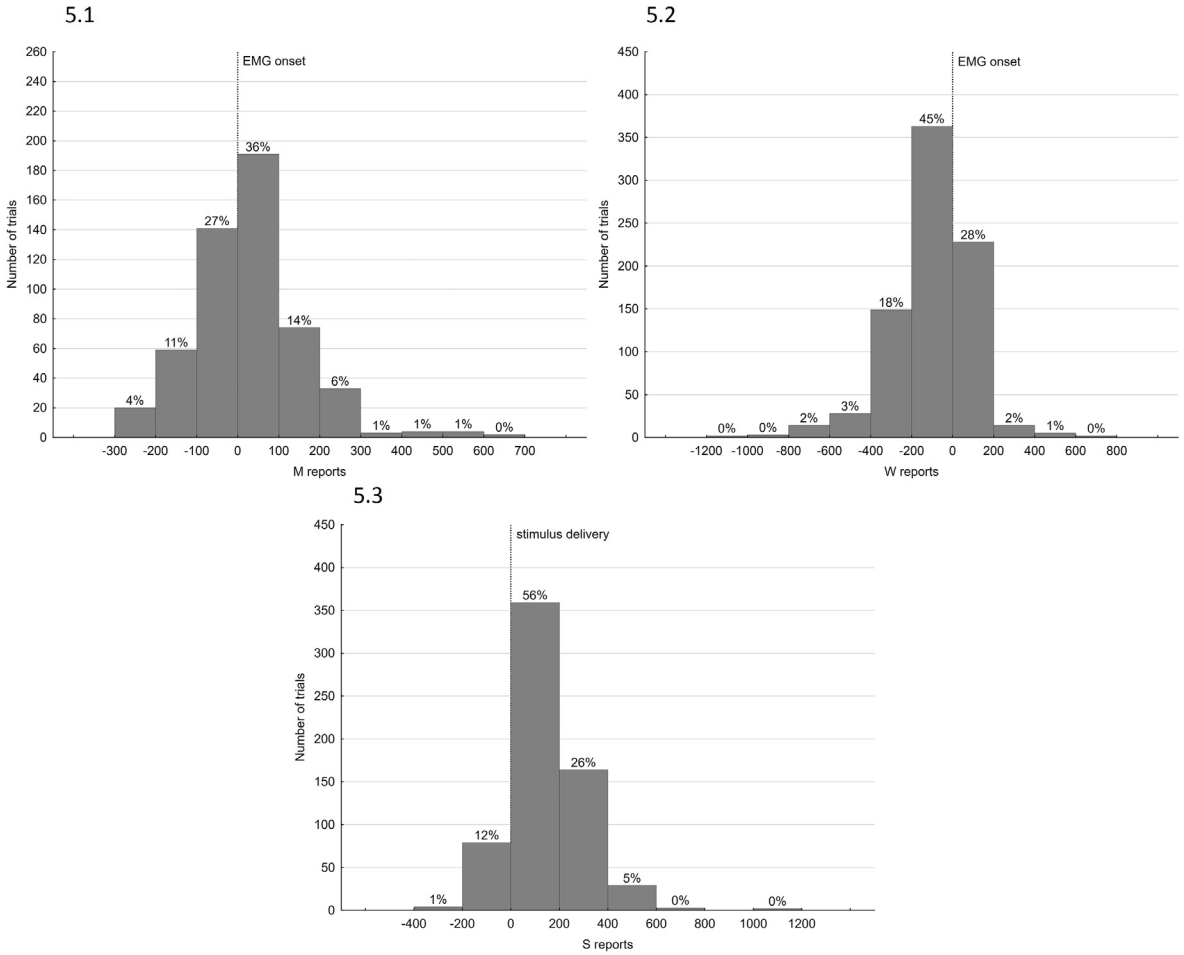

Fig. 5. The histograms of the M(A) reports (n = 531, Fig. 4.1) and W(A) reports (n = 808, Fig. 4.2) compared to the EMG onset and S(A) reports (n = 640, Fig. 4.3) compared to the skin stimulus delivery.

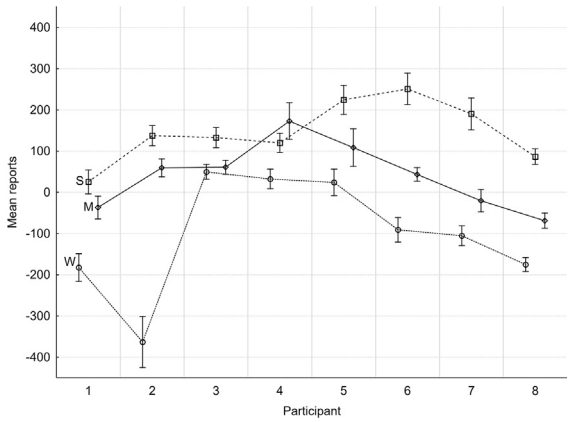

Fig. 6. A graphical representation of the reports' means based on individual participants (x-axis) and the series type (separate lines). Note that the participants' order is arbitrary and that the slope of the connecting lines does not provide any information. The vertical bars denote the 95% confidence intervals. Value 0 on the y-axis denotes the EMG onset or the skin stimulus delivery.

The graph suggests two intriguing findings regarding the W reports: first, that three participants reported the W values on average after the EMG onset, and second, that participant 3 tended to state similar reports regardless of the W and M instructions. These findings are further discussed in Section 4.4.

It is reasonable to ask whether the absolute reports differ based on the series type (M, W, S). To answer the question, we employed

Table 3
An overview of individual effects in the mixed-effect model analyzing the M, W and S reports collected within the absolute mode of recall. The random factors are included, but not displayed. The symbol ** signifies $p < 0.01$, *** signifies $p < 0.001$.

Factor	Estimate	Std. error	*t* value (df)	p value
Intercept	41.13	25.20	1.632 (7.90)	0.142
Series (S)	106.09	26.90	3.944 (7.80)	0.004**
Series (W)	−146.49	37.23	−3.934 (8.00)	0.004**
Session	9.26	2.21	4.194 (1950.80)	< 0.001***
Trial	0.81	0.24	3.316 (1940.90)	< 0.001***

a mixed-effect linear model testing the effect of the series types (**Series**) on the reported value, with following covariates: **Participant** (a random categorical factor describing the differences between the individual participants), **Trial** (a fixed continuous factor describing the succession of individual trials in the series) and **Session** (a fixed continuous factor describing the differences between individual sessions). The Trial and Session factors were centered by subtracting the mean. The covariates are all included mostly to consider the fact that the individual trials are not independent (every 40 trials in a single series came from the same participant, series of the same type were always parts of different sessions and the trials might have brought different reports based on their succession in the respective series). We also included the participant-series interaction term (**Participant*Series**; random factor) as a covariate to take into account the possibility that the participants may provide different reports in specific series. We excluded 15 trials as residual outliers. The model exhibits a satisfying fit and explains 54.8% of the reports' variability (30.8% if the random factors Participant and Participant*Series are excluded). For an overview of the effects, see Table 3.

The main effect of the factor Series is highly significant [$X^2(2) = 13.804$, $p = 0.001$]. These results suggest that there is an overall difference between the M, W and S reports – specifically, the W reports are significantly earlier than the M reports [t (8.00) = 3.934p = .004], the M reports are significantly earlier than the S reports [t(7.80) = 3.944, p = .004] and the W reports are significantly earlier than the S reports [t(8.00) = 5.968, p < .001]. For the estimates comparison of the M(A), W(A) and S(A) reports, see Fig. 7.

Additionally, we tested whether the W reports change when the W task is carried out before or after the M task (see Libet et al., 1983, p. 632; Dominik et al., 2017). We analyzed the **Order** effect (W-M or M-W) on the W reports (**Trial** included as a fixed covariate, combination of **Participant** and **Session** as a random covariate). We found that the Order has no significant effect on the W reports [t(16) = 0.162, p = 0.873], although the W reports acquired after the M task were indeed earlier (M = −115.9, 95% CI [−224.4, −7.4]) than the W reports acquired before the M task (M = −103.9, 95% CI [−212.3, 4.4]).

3.2.2. Order introspective reports

The reports made in the trials with the order (O) mode of recall can be calculated as means for the whole series but not the individual trials. Therefore, we cannot calculate the SD for the series, but we can calculate the SEM (standard error of the mean) as the standard deviation of the mean estimates. Based on that calculation, we can also assess the 95% confidence intervals for the mean based on the SEM and the number of series (not the number of trials as in the case of the A reports).

The mean order report of the movement initiation, i.e. the M(O) report, was remarkably close to the EMG onset (M = 0.0, SEM = 83.5). The mean W(O) report preceded the EMG onset by about 70 ms (M = −70.3, SEM = 104.7). The S(O) reports followed the skin stimulus slightly (M = 34.1, SEM = 69.4). Because the observed values are means and not individual reports, we cannot conduct the same regression procedure as in the case of the absolute reports and make the estimates of the M(O), W(O) and S(O) means after taking the participants, session and trial into account. For an overview of the actual means and the confidence intervals, see Fig. 8.

Finally, we also present the comparison of the A and O modes of recall. Because the Figs. 7 and 8 present different approaches to the mean estimation, we decided to calculate the means and the confidence intervals for the A mode of recall the same way as we did in the case of the O reports (i.e. for the series means, not the individual trials). The resulting estimates are presented in Fig. 9. The figure shows that the order reports tend to be less extreme (i.e. closer to the EMG onset or the stimulus delivery). Inspection of the 95% confidence intervals suggests that there is a significant difference between the S reports made in the opposite modes of recall; other types of reports do not differ significantly based on the mode of recall.

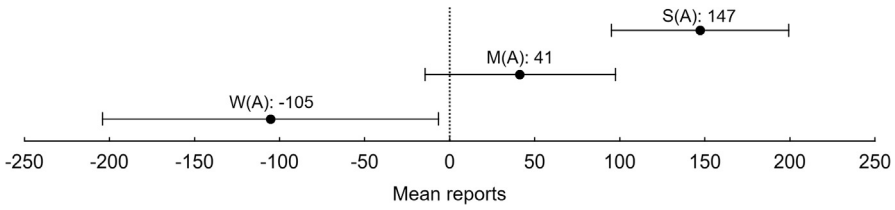

Fig. 7. A comparison of the estimates of the W(A), M(A) and S(A) reports. The horizontal bars denote the 95% confidence intervals. Value 0 on the x-axis denotes the EMG onset for the M and W reports and skin stimulus delivery for the S reports. Note that these results do not correspond precisely to the means for the absolute reports presented at the beginning of Section 3.2.1 due to the inclusion of covariates' effects in the present estimation.

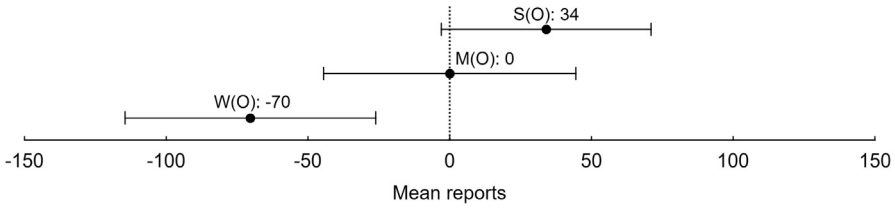

Fig. 8. A comparison of the mean W(O), M(O) and S(O) reports. The horizontal bars denote the 95% confidence intervals. Value 0 on the x-axis denotes the EMG onset for the M and W reports and skin stimulus delivery for the S reports.

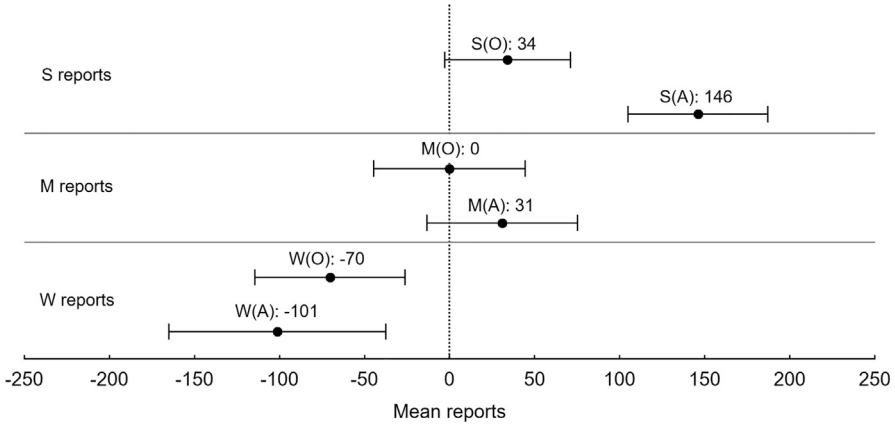

Fig. 9. A comparison of the mean W, M and S reports made in the absolute (A) and order (O) modes of recall. The horizontal bars denote the 95% confidence intervals. Value 0 on the y-axis denotes the EMG onset for the M and W reports and skin stimulus delivery for the S reports.

3.3. Event-related potentials occurrence and timing

We examined two types of event-related potentials (ERPs for short) – the readiness-potential (RP) and the P300 wave (see Section 2.6). The RP is expected to precede the movement onset in the M, W, P and Pv series and their presence and onsets were assessed using two methods (the eye-ball MN method and the computational $RP_{90\%}$ method). The P300 is a cognitive ERP expected to follow the stimulus presentation in the S series.

Contrary to our original expectations, the EEG recordings contained a large amount of noise, which was true even after averaging the signal. Even though this may limit the validity of our findings, our data still demonstrate many interesting points.

Most importantly, the grand averages in the corresponding series types (i.e. the M/W, P, Pv and S) contain the expected waveforms distinctly identifiable on the C_z, C_3, C_4 and P_3 electrodes (see Fig. 10).[7] The Fp_1 and Fp_2 electrodes recordings were often distorted and did not exhibit the expected potentials; these were thus excluded from the grand average plots. Table 4 shows how many trials were included into each grand average calculation.

Fig. 10.1 shows the general readiness potential negativity with the beginning preceding the EMG onset substantially. The grand-averaged RP has the largest amplitude on the C_z, followed by the C_3 (contralateral to the moving hand) and C_4 (ipsilateral to the moving hand). The grand-averaged RP on the P_3 electrode is irregular and exhibits a positive spike near the negativity peak. The grand-averaged RP on the C_z resembles a ramp-like shape characteristic for the type I RP, but it begins later than the type I RP is expected to. On the C_3 and C_4, the grand-averaged RP resembles a dome-shaped type II RP.

Fig. 10.2 demonstrates clearly two important points. First, no negativity with the RP characteristics precedes the stimulus onset. Second, the stimulus delivery is followed by a wave complex ending in a large positive spike, which we identified as the P300 positivity.

Fig. 10.3 shows that when a subject is instructed to make the movement at the target time, the movement is preceded by large negativity preceding the movement by about 1500 ms. This negative potential does not exhibit typical characteristics for any of the RP types – it is dome-shaped, but begins more than 700 ms before the movement, so it should not be regarded type II RP; it also does not fit the pre-set RP description by Libet et al. (1982, p. 330), who suggested that the pre-set RP should exhibit a ramp-like form. Therefore, as stated in Section 2.6, we decided to extend the pre-set RP definition to any early onset negativity preceding the movement during the P and Pv series.

Fig. 10.4 shows a negativity more closely fitting Libet's definition of the pre-set RP. The potential precedes the target time (and

[7] All our ERP plots depict the positive charge above the baseline. We are aware that this disagrees with some older conventions, but we find it easier to communicate the results this way when reporting both negative and positive potentials.

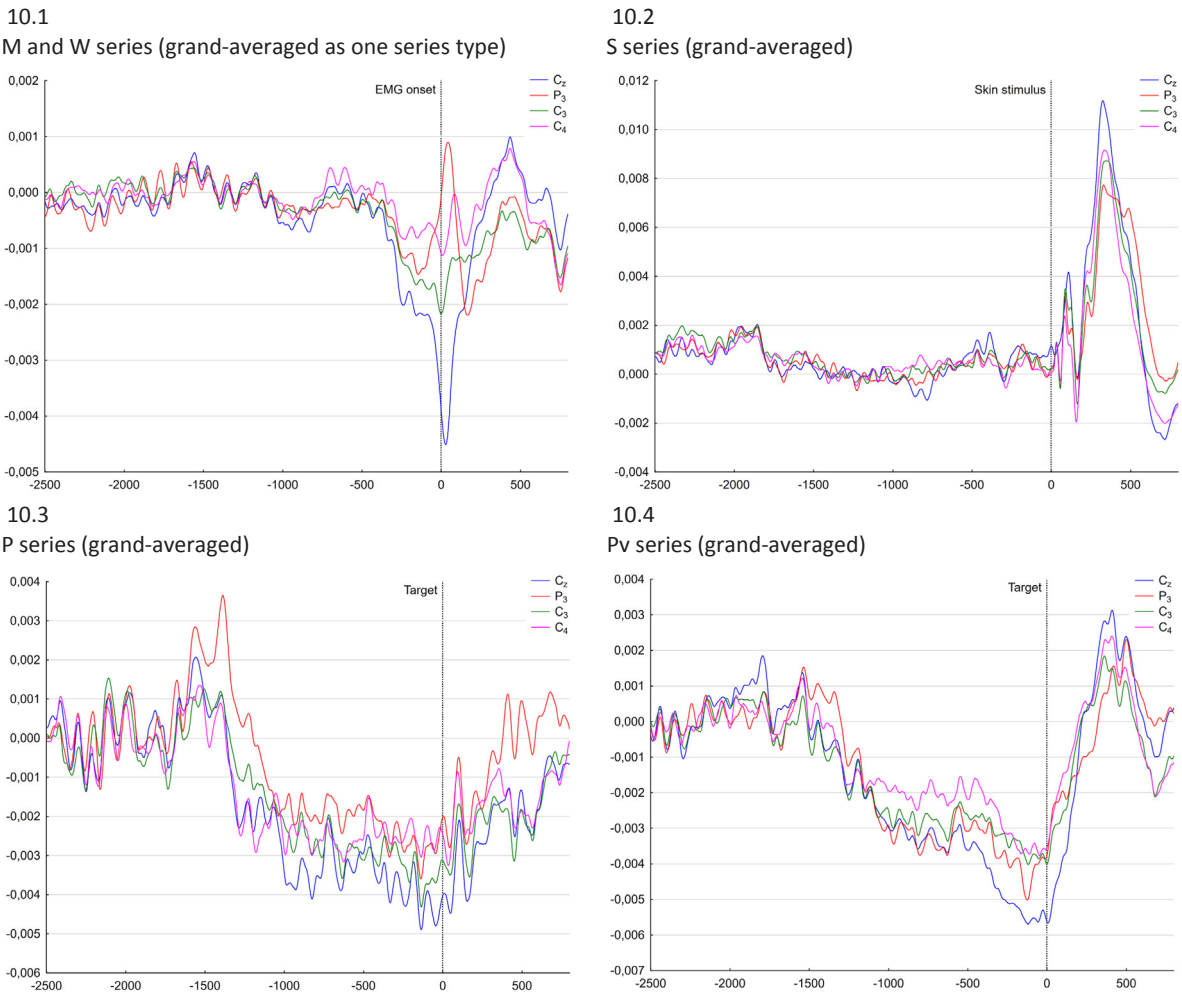

Fig. 10. The grand-averaged ERPs in the four series types. Fig. 10.1 shows distinct RP negativity beginning prior to the EMG onset. Fig. 10.2 shows a P300 positivity following the stimulus. Fig. 10.3 shows irregular dome-shaped negativity beginning long before the pre-set target point on Libet's clock was reached in the P series. Fig. 10.4 demonstrates ramp-like negativity beginning long before the target point was reached in the Pv series, which in this case should be the time when the vetoed movement was supposed to be realized. The baseline in Fig. 10.1 and 10.2 was estimated as the mean voltage in the time interval $< -1500;\ -1000 >$; the baseline in Fig. 10.3 and 10.4 is estimated as the mean voltage in the time interval $< -2500;\ -2000 >$ (because the observed potentials in the P and Pv series are longer).

Table 4
Number of trials included into each grand-averaged ERP plot. The M and W trials are in this case considered to be equivalent.

	C_z	P_3	C_3	C_4
M/W	1759	1818	1672	1747
S	464	835	776	792
P	197	197	194	198
Pv	498	497	476	469

the moment of conscious veto, respectively) by about 1500 ms and has a ramp-like growth.

These results show that although the recordings were noisy, the outcomes seem to be valid. We will now investigate the individual series types and the respective ERPs found in them. Table 5 lists which RP types were identified on C_z in which series (we consider the C_z electrode to be the most informative). We only list the series in which a RP on the C_z electrode was found using the MN method. The $RP_{90\%}$ method seemed to be more sensitive to RPs obscured by the noise hindering the MN detection; however, as we aimed to classify all RPs as type I, II, III or pre-set, we used the $RP_{90\%}$ as a supplementary method, because if the RP was not identified using the MN method, the RP type could not have been assessed. Because of the low frequency of identified RPs, we cannot perform any inferential statistical procedure and will thus report descriptive statistics only.

Table 5
Number of ERPs found on the C_z electrode in individual series. The bold frequencies mark which ERPs are expected to occur in the respective series. The percentages show the relative frequency of the respective RP type compared to the number of valid series of the given kind. The percentages are not stated in the P300 column, because the P300 did not in our procedure rule out the RP occurrence (the percentage would therefore be confusing as their row sum would exceed 100%).

	RP I	RP II	RP III	pre-set RP	no RP found	P300	valid series
M	**3 (15%)**	**4 (20%)**	**2 (10%)**	0 (0%)	11 (55%)	0	20
W	**6 (19%)**	**6 (19%)**	**4 (13%)**	0 (0%)	16 (50%)	4	32
P	1 (8%)	2 (15%)	0 (0%)	**7 (54%)**	3 (23%)	2	13
Pv	0 (0%)	0 (0%)	0 (0%)	**6 (46%)**	7 (54%)	0	13
S	0 (0%)	0 (0%)	0 (0%)	1 (8%)	**11 (92%)**	**8**	12

3.3.1. ERPs in the M and W series

Altogether, we analyzed 303 plots with valid EEG acquired in the M or W series, including all 6 electrodes. We found **21 ramp-like type I RPs** in 14 series (9 on C_z, 2 on C_3, 5 on C_4, 3 on P_3, 1 on Fp_1 and 1 on Fp_2) across 11 sessions. The average type I RPs' onset was estimated to be −667.4 ms (SD = 168.7) using the MN method and −580.3 ms (SD = 185.3) using the $RP_{90\%}$ method. Fig. 11.1 demonstrates typical type I RP.

Further, we found **37 dome-shaped type II RPs** in 18 series (10 on C_z, 9 on C_3, 5 on C_4, 9 on P_3, 1 on Fp_1 and 3 on Fp_2) across 13 sessions. The average type II RPs' onset was estimated to be −377.3 ms (SD = 142.4) using the MN method and −421.1 ms (SD = 223.4) using the $RP_{90\%}$. Fig. 11.2 demonstrates typical type II RP.

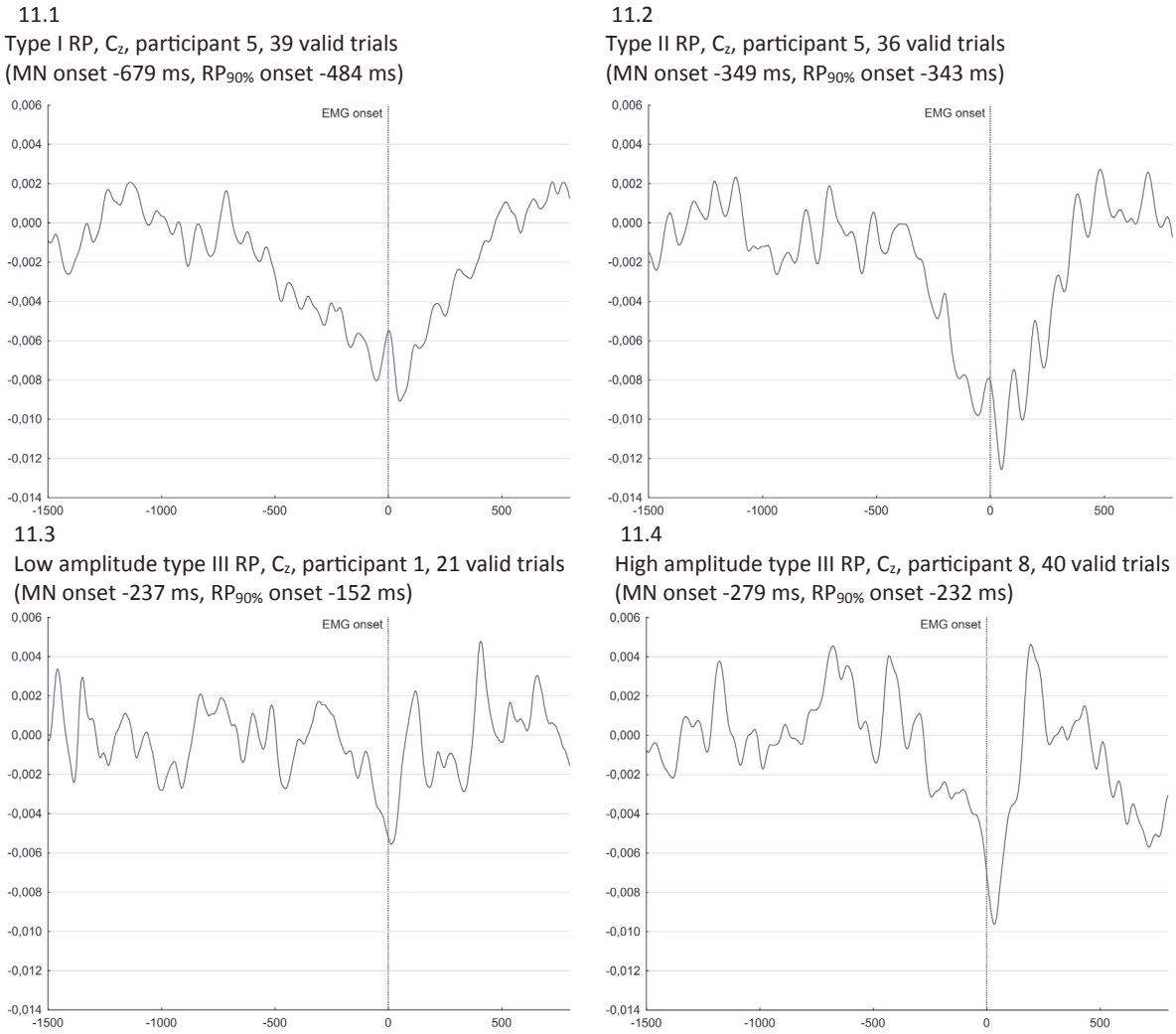

Fig. 11. Examples of type I, type II and type III RPs on the C_z electrode. The plots depict averaged trials in one series.

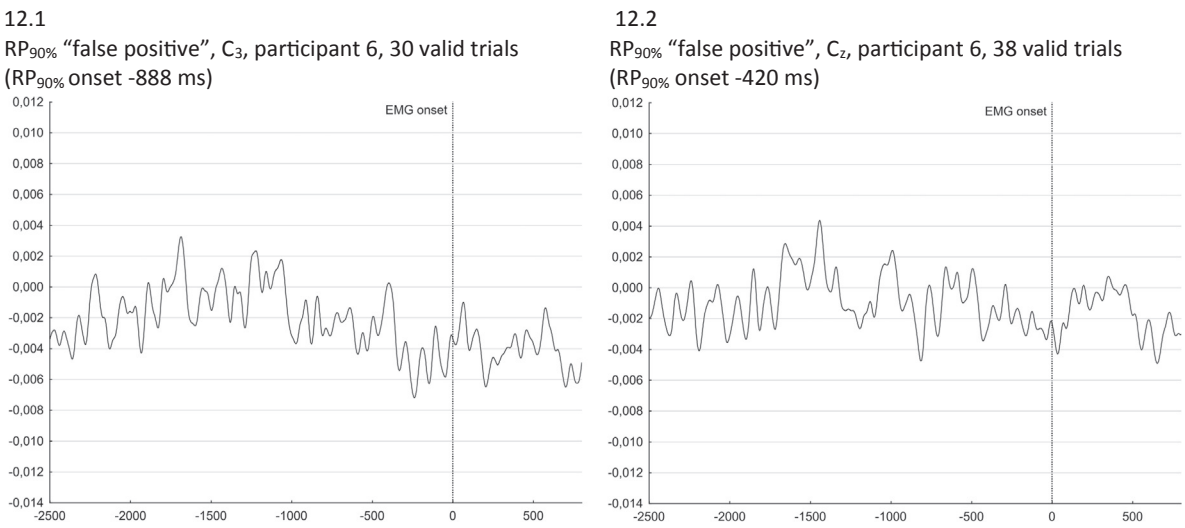

Fig. 12. Examples of "false positive" results using the $RP_{90\%}$ method. In this context, we consider the eye-ball detection to be the "true positive".

The type III RPs (negative potentials with late onset) were found to be almost as common as type I RPs – we found **19 type III RPs** in 11 series (6 on C_z, 4 on C_3, 3 on C_4, 2 on P_3, 2 on Fp_1 and 2 on Fp_2) across 10 sessions. The average onset was −250.7 ms (SD = 84.2) using the MN method and −249.7 ms (SD = 201.9) using the $RP_{90\%}$ method. We found that the type III RPs can be further divided into low amplitude RP (RP with late onset and low voltage on its peak) and high amplitude RP (RP with late onset, sharp shape and high voltage on its peak). Fig. 11.3 demonstrates typical low amplitude type III RP, Fig. 11.4 depicts typical high amplitude type III RP.

We investigated the differences between the MN and $RP_{90\%}$ methods of RP estimation in the W and M series. On average, the MN method estimated earlier RP onsets by mere 3.0 ms (SD = 224.1). The correlation between the methods' estimates is moderate [r = 0.50]. The $RP_{90\%}$ method recognized every RP identified in the MN estimation. The number of the opposite cases is, however, substantial – of the 226 plots in which no RP was found using the MN method, the $RP_{90\%}$ identified 130 RPs (57.5%). See Fig. 12.1 and 12.2 for examples of these $RP_{90\%}$ "false positive" results.[8]

3.3.2. ERPs in the P and Pv series

In both the P and Pv series, a specific type of RP with extremely early onset was expected. Out of the 75 valid EEG plots from the **P series** we found **2 type I RPs** (in 2 series), **14 type II RPs** (in 5 series), **1 type III RP** (low amplitude) and **25 RPs identified as the pre-set RPs** (found in 9 series; the pre-set RPs most closely resembled the irregular RP shape depicted in Fig. 10.3). Therefore, type II and pre-set RPs are the predominant RP types present in the P series. The type II RPs' onset was on average −406.8 ms (SD = 127.0) using the MN method and −376.4 ms (SD = 166.3) using the $RP_{90\%}$ method. These type II RPs did not exhibit any notable differences from the type II RPs acquired in the M and W series (for an example, see Fig. 13.1, compare to Fig. 11.2). The pre-set RPs' onset was on average −1267.4 ms (SD = 185.9) using the MN method and −1035.6 ms (SD = 144.5) using the $RP_{90\%}$ method (one of the 25 pre-set RPs identified using the MN method was undetected by the $RP_{90\%}$). A striking characteristic of the pre-set RPs in the P series was that they tended to reach the peak negativity long before the target time (see Fig. 13.2). Another interesting fact is that not a single pre-set RP was ever identified on the Fp_1 and Fp_2 electrodes (in neither of the P and Pv series); in the P series, 7 pre-set RPs were detected on C_z, 6 on C_3, 4 on C_4 and 8 on P_3.

In the **Pv series**, we identified **2 type III RPs** and **18 pre-set RPs** (in 7 series) out of 75 valid EEG plots. The mean pre-set RP onset in the Pv series was −1192.1 ms (SD = 180.2) using the MN method and −989.6 ms (SD = 492.6) using the $RP_{90\%}$ method. There are some notable differences between the pre-set RPs in the P and in the Pv series. First, even though Fig. 10.4 demonstrates a potential similar to type I RP, the averaged plots of most individual Pv series from individual electrodes suggest that the vetoed pre-set RP tended to fully return to baseline before the target time was reached (see Fig. 13.3), sometime by even more than 500 ms. This might be the reason why the $RP_{90\%}$ method did not detect 2 of the 20 pre-set RPs detected by the MN method. Second, the veto pre-set RPs tended to be preceded by a short positivity in some series; the following RP usually had lower amplitude (see Fig. 13.4).

3.3.3. Erps in the S series

The EEG recordings in the S series are not supposed to contain any RP negativities, but P300 positivity following the stimulus delivery is expected. Nevertheless, the examiners detected 8 RPs across 114 valid EEG plots in the S series (3 type II RPs, 2 type III RP, 3 pre-set RPs; see Section 4.5).

The **P300 wave was detected in 59 of the 114 plots**, in 18 separate series. The four examiners analyzing the P300 mostly agreed

[8] In this context, we call the result "false positive" if the $RP_{90\%}$ detects an RP where RP_{MN} does not.

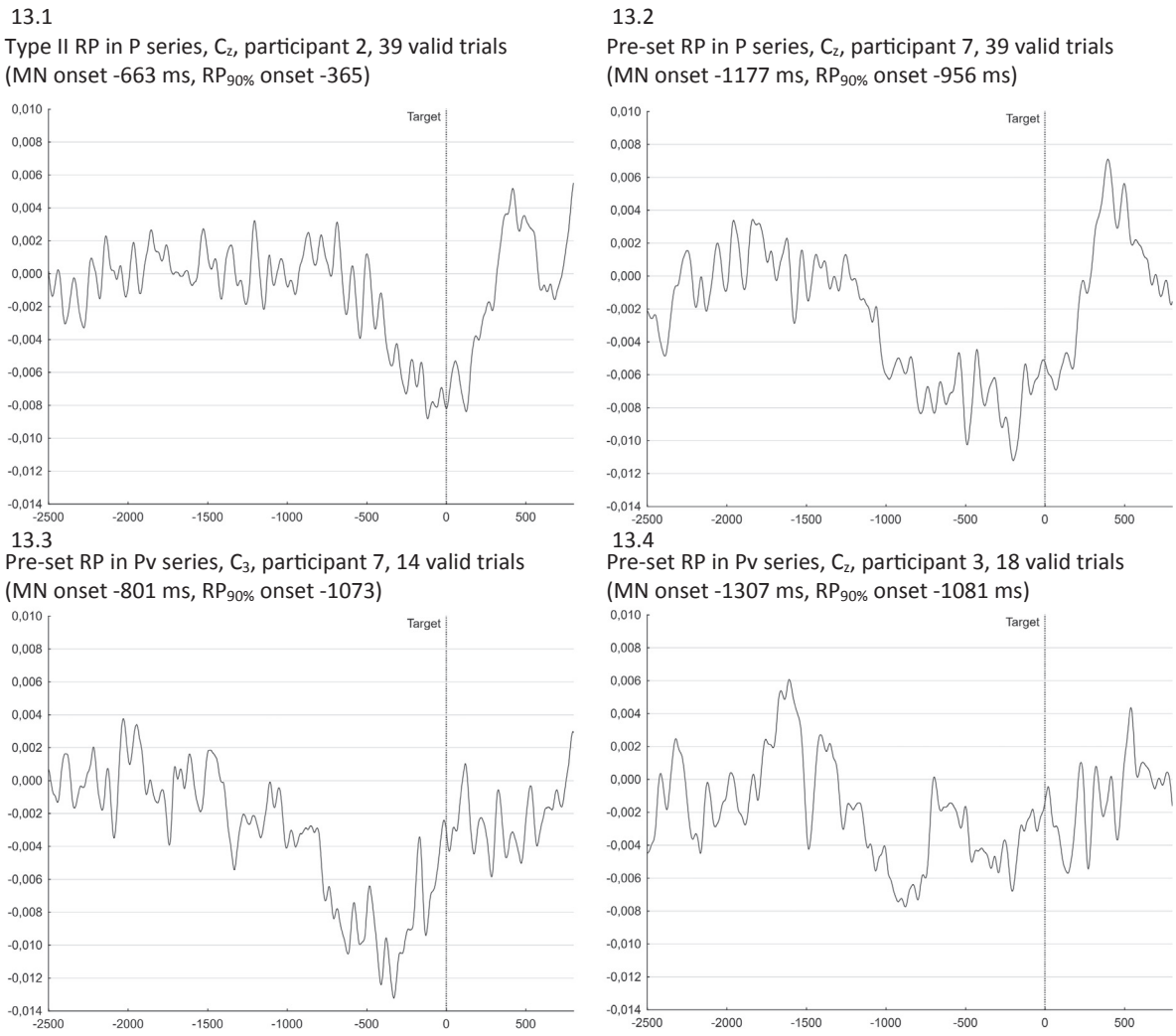

Fig. 13. Examples of different RP types identified in the P and Pv series.

on the P300 wave detection – 30 waves were detected by all four examiners, 14 waves were detected by three examiners, 7 by two and 7 by only one examiner (the waves detected by two and fewer examiners must have been confirmed with special care by the researcher who merged the assessments into a single conclusion). The P300 waves in the S series were mostly found on the C_3 and C_4 electrodes (15 occurrences on C_3, 16 on C_4); 13 waves were found on P_3, 8 on C_z, 3 on Fp_1 and 4 on Fp_2. For an example of a detected P300 waveform, see Fig. 14.1.

The examiners also found 16 P300 waveforms in other than the S series, namely in 10 plots within five W series and in 6 plots within three P series. These P300 waveforms might have in some cases been misjudged rebounds of preceding RPs, because 10 of the 16 waves were detected together with an RP (for an example, see Fig. 14.2).

3.4. Interindividual differences

It has been suggested in Section 3.2.1 that the participants differed from each other in their introspective reports. In this section, we would like to summarize most of the available data into a comprehensive Table 6, which can be compared to Libet's results table (see Libet et al., 1983, p. 630, Table 1).

Table 6 shows some important points, which are commented on in Section 4.4. First, some of the mean W reports referred to time after the EMG onset, especially in the cases of participants 3 and 4. Second, the introspective reports generally differed substantially between the participants, as was already suggested in Section 3.2.1. Third, the noisiness of the EEG recordings led to a rejection of many RP measurements. Furthermore, many of the remaining RP plots contained no RP detected by the MN method and in some cases neither by the MN and $RP_{90\%}$ method.

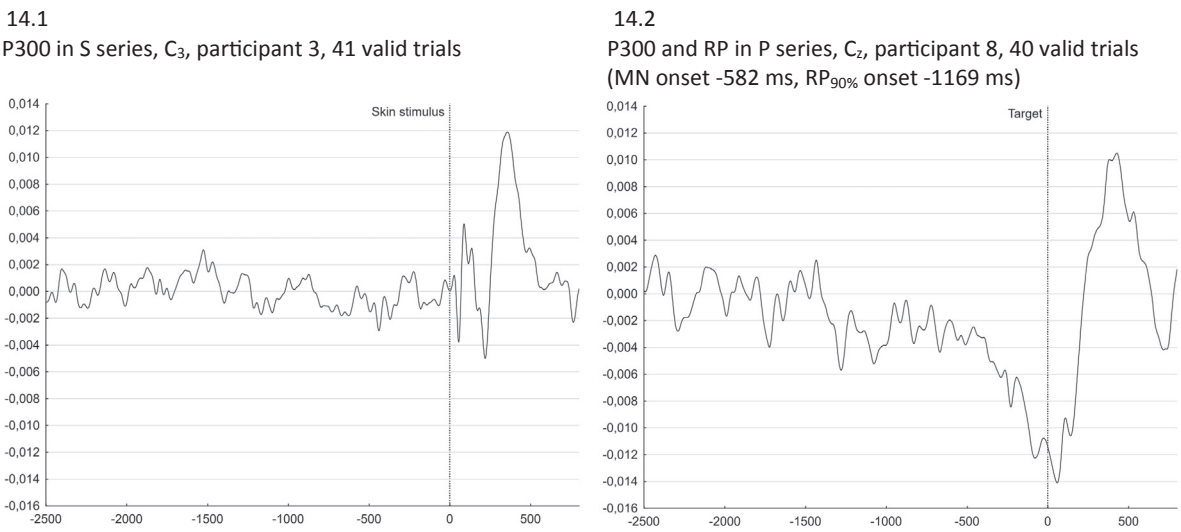

Fig. 14. Examples of the P300 waves.

3.5. Series-specific effects

In this brief section, we report how the participants answered our control questions and what possible effects could potential exceptionalities have on other results. The two questions are stated in Section 2.5 and always pertained W and M series. **The first question** asked whether the participant felt that **the movement was pre-planned** and not spontaneous. In most sessions, the participants stated that they did not feel any pre-planning before moving the finger. If the participants answered that they did (which happened in 17 of the 48 sessions), they usually stated that the pre-planning occurred in 1 to 4 trials. In two cases, the reported frequencies of the pre-planning were higher than 4. In the first case, participant 8 reported that approximately 8 trials in one W(A) series were pre-planned – the ERP detected in this series on C_3, C_4 and P_3 was type II RP with MN onsets -699 ms, -699 ms and -898 ms, respectively; the W(A) reports in the series were earlier than the sample average ($M = -212.4$ ms, $SD = 100.3$). In the second case, the same participant reported approximately 6 trials in one M(A) series to be pre-planned – this time, the RP detected on C_z, C_3, C_4 and P_3 was type II with MN onsets -309 ms, -315 ms, -302 ms and -298 ms, respectively; the M(A) reports in this series were notably earlier than the sample average ($M = -85.0$ ms, $SD = 78.9$).

The second question asked whether the participant **was surprised by his or her own movement**. In 30 of the 48 sessions, the participants stated that they were surprised by the movement in at least one trial. Reported frequencies of this happening were overall similar to the answers to the first question – participants usually reported approximately 4 movements in a series to be surprising. Participant 1 reported six times that the situation happened in 10 and more trials in a single series. We examined the results from these six series (3 W(A), 1 M(A), 1 W(O) and 1 M(O) series). We inspected 36 corresponding EEG plots and found that the RPs were rarely detected using the MN method. If any RPs were found (in 9 cases), they were predominantly type II and type III with low amplitude and late onset (with an exception of type I RPs found on the Fp_1 and Fp_2 electrodes in one of the W(A) series). The introspective reports in said series were as follows: mean M(A) report was 8.1 ms ($SD = 99.9$), mean W(A) report was -182 ms ($SD = 169.9$), M(O) mean estimate was 46.9 ms and W(O) mean estimate was -48.9 ms; these results do not seem extraordinary in any way among other participants' introspective reports.

4. Discussion

This paper aims to replicate the outcomes of Libet et al. (1982, 1983) and Libet (1985) by conducting a study designed to follow the original methodology as closely as possible. The experiment is complex and as our data suggest, its outcomes are substantially dependent on data collection and analysis procedures. In the following six sections, we discuss technical issues and effects of our methodological choices, as well as the implications of our replication and other recent empirical studies for the validity of Libet's experiment.

4.1. Procedure and instruction discussions

Our replication differs from Libet's methodology in several ways. One of them is the sample – we tested eight participants instead of six (Libet et al., 1982, p. 323; Libet et al., 1983, p. 624); this may seem as an irrelevant change, but for future Libet-style experiments analyzing both the W and M series in both the A and O modes of recall, we strongly recommend using sample sizes divisible by 4 to allow for complete experimental conditions rotation (see Section 2.5, Table 2). We also strongly recommend conducting a full-scale replication with a much larger sample and attempt to find the factors responsible for the large interindividual

Table 6
Overall summary of the data obtained in the W, M and S series. Session 1 is not displayed, as it served training purposes only. The W and M reports are calculated in relation to the EMG onset, S reports to the skin stimulus delivery; their SDs are calculated for the absolute mode of recall only; in the order mode of recall, only mean estimates are presented. In the RP onsets section, both RP_{MN} and $RP_{90\%}$ on C_z are presented, but the RP type is stated only when the RP was detected using the eye-ball MN method. Label "low EEG quality" signifies that C_z electrode collected excessive noise, so the signal could not be analyzed; "no RP" means that the RP was not detected; "not enough valid trials" means that the averaged EEG plot was calculated from fewer than 10 trials.

Participant	Session	Mode of recall	Introspective reports (ms)									RP onsets on Cz					
			W			M			S			RP in W series			RP in M series		
			M	SD	n	M	SD	n	M	SD	n	Type	RP_{MN}	$RP_{90\%}$	Type	RP_{MN}	$RP_{90\%}$
1	2	A	−104	179	38	−79	127	40	+3	148	40		no RP	−33		no RP	−2500
	3	O	−19		41	−14		41	+5		41		no RP	no RP		no RP	no RP
	4	A	−151	89	37	+8	100	38	+48	106	40	II	−216	−742	II	−346	−270
	5	O	−49		41	+47		41	+43		41		no RP	no RP		no RP	−14
	6	A	−342	140	26								no RP	−470			
	7	O	−131		41							III	−237	−152			
2	2	O	−205		41	−70		41	−3		41	I	−735	−589		no RP	−118
	3	A	−512	287	34	+85	92	29	+127	118	40	I	−517	−418		no RP	−888
	4	O	−256		41	−58		41	+43		41	not enough valid trials				no RP	−112
	5	A	−277	246	21	+32	58	27	+148	102	40	II	−272	−409	II	−369	−324
	6	O	−183		41							II	−433	−334			
	7	A	−233	190	25								no RP	−795			
3	2	A	+32	83	25	+44	73	25	+134	113	40		no RP	−401		no RP	no RP
	3	O	+73		41	+54		41	+28		41	low EEG quality			low EEG quality		
	4	A	+43	86	34	+73	56	33	+131	107	40	low EEG quality			low EEG quality		
	5	O	+80		41	+71		41	+58		41	III	−393	−331	II	−309	−477
	6	A	+79	70	23								no RP	−354			
	7	O	−3		41								no RP	−803			
4	2	O	−41		41	+47		41	+58		41	low EEG quality			low EEG quality		
	3	A	+13	116	38	+173	135	38	+121	121	40	low EEG quality			low EEG quality		
	4	O	+70		41	+69		41	+58		41	low EEG quality			low EEG quality		
	5	A	+16	126	33				+120	90	40	low EEG quality			low EEG quality		
	6	O	+54		41							low EEG quality					
	7	A	+76	110	30								no RP	−2			
5	2	A	+11	234	39	+126	224	31	+236	156	40	II	−349	−343	I	−582	−937
	3	O	−46		41	−3		41	+50		41	I	−789	−444	I	−504	−859
	4	A	+59	119	39	+95	161	39	+212	162	40	II	−391	−270	I	−679	−484
	5	O	+4		41	+5		41	+88		41	I	−858	−718	II	−420	−337
	6	A	−2	122	32							low EEG quality					
	7	O	−37		41							low EEG quality					
6	2	O	−71		41	−93		41	−93		41	low EEG quality			low EEG quality		
	3	A	−44	99	34	+71	61	37	+230	118	40	low EEG quality			low EEG quality		
	4	O	−246		41	−147		41	−153		41	low EEG quality			low EEG quality		
	5	A	−257	129	30	+14	68	35	+272	212	40		no RP	no RP		no RP	−360
	6	O	−127		41								no RP	−71			
	7	A	0	84	37								no RP	−420			
7	2	A	−195	143	39	−94	89	40	+117	189	40	low EEG quality			low EEG quality		
	3	O	−59		41	+21		41	+88		41	low EEG quality			low EEG quality		
	4	A	−80	78	39	+54	104	40	+263	118	40	low EEG quality			low EEG quality		
	5	O	−93		41	+126		41	+110		41	II	−352	−279		no RP	−305
	6	A	−34	94	35							I	−1037	−864			
	7	O	+90		41								no RP	−510			
8	2	O	−124		41	+98		41	+103		41		no RP	−34	III	−270	−27
	3	A	−149	93	40	−52	83	39	+99	81	40	III	−279	−232	III	−147	−90
	4	O	−195		41	−152		41	+65		41		no RP	no RP		no RP	−3
	5	A	−164	74	40	−85	79	40	+74	88	40		no RP	no RP		no RP	−22
	6	O	−172		41							III	−334	−194			
	7	A	−212	100	40							I	−563	−451			

variability. The reason is that our results and results of other researchers, Libet including, suggest that some participants tend to differ considerably in their introspective reports from the rest of the sample (e.g. later W reports of participants 3 and 4 in our study, see Table 6; later M and S reports of participant B.D. in Libet et al., 1983; later W reports of participant IB in Keller & Heckhausen, 1990; earlier W reports of participant 5 in Verbaarschot et al., 2015). Simple grand average results from the introspective data are therefore meaningless, because they do not account for the interindividual differences (that is also why we used the mixed-effect linear model to analyze the W, M and S reports, see Section 3.2.1).

Another problematic element of the experimental procedure is the instruction, especially when asking for the introspective report, which we signify as W in this text. In Libet's original experiment, the participant was instructed to "*report the time of appearance of his conscious awareness of 'wanting' to perform a given self-initiated movement*" (Libet et al., 1983, p. 627), but Libet also states that the participants described the experience as an "urge", "intention" or "decision". Soon et al. (2008, p. 543) instructed the participants to "*press a button as soon as they felt the urge to do so*", but to report "*when their motor decision was consciously made*".[9] Trevena and Miller (2002, p. 172) asked the participants to report the dot position "*at the time of the decision to 'go now'*". Verbaarschot et al. (2015, p. 301) asked the participants *"to report the onset of their intention to act"*. Caspar and Cleeremans (2016, p. 4) instructed the participants "*to report the location that the black spot occupied at the time they had first decided to press the key*". Most of these studies diverged notably in the acquired W results which suggests that the instruction formulation might influence the results significantly. In our case the instruction was to report the clock time when the participant "*realized the first urge to press the mouse button*".[10] We decided to use the word "urge" for following reasons. First, Libet et al. (1983, p. 627) stated that *"subjects usually settled for the words 'wanting' or 'urge'"*; we intended to use one of these and, as "wanting" does not sound natural in Czech sentence, we decided to choose the "urge" option. Second, Pockett & Purdy (2011) argue that there is a difference in the experiment results when instructions contain either "decision" or "urge" and that the "urge" instruction replicates Libet's results.

Another potentially significant deviation from Libet's methodology is that we did not provide any feedback to the participants based on their S reports. Our reasons were explained at the end of Section 2.4. As Gomes (1998, p. 590) suggests, providing feedback on how accurate the participant is in every S series throughout the experiment will certainly change the results leading to large inconsistencies in data obtained in the first few sessions compared to data acquired in the last sessions. Furthermore, based on the results of Danquah et al. (2008), we can assume that the feedback training would have varying effects on the M and W reports based on the stimulus modality in the S series. Based on these obvious problems, we decided not to tell the participants how accurate they were in the S series, nor to perform any S trainings at the beginning of the sessions (as these would naturally require us to provide the feedback to have any training effect). One could argue that the fact that we did not provide the feedback was the cause for the strikingly late S reports in the S(A) series. However strange it may appear that this would influence only the S(A) reports and not the S (O) reports, we admit that the S series might indeed require training to be valid. Based on this intricacy, we also later refrained from correcting the W and M reports using the S reports, as Libet suggested (Libet et al., 1983, p. 631).

4.2. Technical equipment discussions

Our technical equipment allowed us to perform some of the experimental procedures more precisely. The display used was a normal LCD computer display, which compared to the cathode-ray oscilloscope (CRO) used in Libet's case (Libet et al., 1983, p. 625) offers better contrast (for that reason we also used black and white colors instead of the grey and bright green displayed by the CRO). Also, we presumably increased the reporting precision in the A mode of recall by allowing our participants to click on the reported location.

We did not make changes to the EEG measurements except replacing the C_c and C_i electrodes with C_3 and C_4 (C_c and C_i locations cannot be measured using the standardized 10–20 EEG cap, see Section 2.3). As it turned out, the eye fixation was not as perfect as Libet et al. (1983, p. 323) asserted – after the EEG analyses, we acknowledged that the EOG measurements would in fact reduce the number of epochs rejected due to the eye-movements. We do recommend future researchers to use the EOG to remove the eye-related artifacts from the EEG recordings.

The S series required a skin stimulator, which in Libet's case was electrical (Libet et al., 1982, pp. 323–324). We decided to use a tactile stimulator to eliminate the risk that an electrical impulse would interfere with the EEG data. It might be objected that the tactile skin stimulator is not as precise as the electrical stimulator, because it needs to mechanically move a bolt touching the participant's skin to deliver the stimulus. We agree that this might be an issue, but we did not find any viable measure to check the stimulator's precision other than our own subjective test. However, to eliminate additional potential error in the timing of the trigger signal (sent through a 3.5 mm audio cable) driving the stimulator, we used a check loop recording the difference in signal input, output and repeated input on the MP150 unit which did not show any latencies whatsoever.

4.3. EMG measurements discussions

For technical reasons addressed in Section 2.4, the movement in our study had to be in the form of a mouse click. As argued earlier, this may allow us to investigate some additional issues. Button press is used in many Libet-style experiments and results of this type of response are expected to be equivalent to Libet's original flexion of the wrist or fingers (Libet et al., 1983, p. 625). As it turned out, the mouse click might have a few issues. First, the click produces an auditory feedback, which might shift the introspective reports, as showed by Banks and Isham (2009), so it requires additional measures (such as the earplugs used by us) to overcome. Second, even if the auditory feedback is eliminated, the mouse still provides a slight haptic feedback when pressed – this might raise the question whether the participants relate the W and M reports to the EMG onset or the mouse click. Third, because the movement is small, it happened multiple times in our experiment that the EMG onset was not registered by the electromyograph if the participant conducted the movement not rapidly enough or if the electrodes were not attached precisely (which would be a smaller

[9] In this case, authors also did not use Libet's clock but letter slideshow presenting one letter every 500 ms.

[10] "(…) v okamžiku, kde jste si uvědomil/uvědomila první nutkání stisknout tlačítko" in Czech.

problem if the movement was a flexion of the whole wrist, because the number of activated muscle fibers in such movement is significantly higher). Therefore, our conjecture stated in Section 2.4 that the mouse click provides a better-bounded movement more suitable for Libet's task seems to be incorrect.

The movement issues are also related to our results pertaining the EMG onsets. Besides the 286 trials containing no detectable EMG onset in the M, W and P series, another 318 trials contained EMG preceding the mouse click by more than 150 ms or following it by any amount of time. These results should be further discussed. The EMG onsets might have been undetected for two main reasons: either the integrated EMG level did not reach sufficient threshold to be detected or early artifacts were present in the recording obscuring the EMG onset. The early EMG onsets (< -150 ms) were most probably caused by early supraliminal muscle twitches, while the late EMG onsets (greater than 0 ms) were found presumably due to insufficiently abrupt movement initiation (causing the baseline to gradually raise, never exceeding the threshold based on its IQR, see Section 2.6). Even though these errors were not common among the 3879 M, W and P trials, they may cause issues when averaging the EEG, which relies on the temporal alignment based on the EMG onsets (their non-detection reduces the number of valid epochs entering the averaging procedure). Seemingly, this problem could have also been worked around if the wrist flexion was used instead of the mouse click; nevertheless, the mouse click also serves as a kind of a control mechanism checking that an early artifact is not detected as the EMG onset. Each of the technical solutions seems to have its advantages and disadvantages.

As stated in Section 3.1.1, we also compared the EMG onset data related to the mouse click to the results of Haggard and Eimer (1999). Our results suggest that absolute majority of the EMG onsets precedes the mouse click by 30–90 ms, but that some outliers are present on both ends of the range spanning from -330 to $+106$ ms. It may be needed to perform a study dedicated solely to this issue.

For all the aforementioned reasons, we **recommend caution when using the mouse click to time the introspective reports**.

4.4. Introspective reports discussions

As stated in Section 4.1, we have doubts that presenting grand averaged introspective reports is meaningful. However, to make our results comparable to those of Libet, we presented raw mean reports (calculated from the series means) for M(A) (M = 30.9 ms, SEM = 79.8), M(O) (M = 0.0, SEM = 83.5), W(A) (M = −101.2, SEM = 151.1), W(O) (M = −70.3, SEM = 104.7), S(A) (M = 146.0, SD = 77.3) and S(O) (M = 34.1, SEM = 69.4).

Libet's mean M report was −86 ms (Libet et al., 1983, p. 631), which is by 117 ms earlier than our M(A) report and by 86 ms earlier than our M(O) report. There are other studies reporting their mean M time, with a large amount of inter-study variability: −89 ms (SD = 118, Haggard & Eimer, 1999), 19.8 ms (SD = 39.0, Sirigu et al., 2004, healthy participants), 5.2 ms (SEM = 18.6, Pirio Richardson et al., 2006, healthy participants), −91 ms (SD = 92, Moretto, Schwingenschuh, Katschnig, Bhatia, & Haggard, 2011), −59.9 ms (SE = 5.3, Caspar & Cleeremans, 2016). Compared to all the presented results, our data seem to be generally shifted towards later values, but the difference is not significant in many cases. We can safely rule out that the difference is due to omission of the S series trainings, because no other studies than the one conducted by Libet included it. It seems that the shift may stem from the instruction differences or methodological measures (such as making the report by clicking into the clock face). However, we consider our results to be more valid than those of Libet, because the M reports of our participants are on average far more precise than those of Libet's subjects.

Libet's W reports were −204 ms on average, which is by 105 ms earlier than our mean W(A) report and by 134 ms earlier than our mean W(O) report. The W reports in our replication seemingly differ from Libet's original results, but 95% confidence interval, derived from our mixed-effect model presented in Section 3.2.1, does include Libet's −204 ms, as well as results found by other researchers (e.g. Trevena & Miller, 2002; Pirio Richardson et al., 2006, healthy participants; Schurger et al., 2012; Verbaarschot et al., 2015; Caspar & Cleeremans, 2016). On the other hand, there are also studies with mean W reports earlier than the lower bound of our CI (e.g. Keller & Heckhausen, 1990; Haggard & Eimer, 1999; Sirigu et al., 2004, healthy participants; Edwards et al., 2011, healthy participants; Moretto et al., 2011, healthy participants). It seems that the W reports are extremely variable. We suggest that these discrepancies are to be expected due to the fact that there is large interindividual variability in the W reports (see Section 3.2.1) and that many studies use small samples. The question is why the reports of our participants differ from Libet's original findings. The reasons may be that we did not train the participants' precision in the S series (which, as stated in the previous paragraph, seems unlikely) or that we provided different instructions than Libet, as Libet did not publish the whole instruction text; neither of these, however, necessarily means that our W reports are invalid – it may simply mean that our participants understood the W concept differently than Libet's participants or the participants in other studies. Additionally, we also have to adress one issue already mentioned in the introductory Section 1 – Trevena and Miller (2002) reported that about 40% of the W reports were found to be later than the movement onset. In our study, we made a similar finding – 249 of all 808 valid W reports (30.8%) were found to follow the EMG onset, while the mean W reports of participants 3 and 4 followed the EMG onset almost exclusively (see Section 3.4, Table 6).

This is in fact connected to a question whether the W reports might be susceptible to bias due to instructions or order of the experimental tasks (Dominik et al., 2017; Pockett & Purdy, 2011). We found that the W reports obtained in sessions in which the W task followed the M task were earlier than the W reports obtained before the M task in the respective session; however, the difference was not statistically significant. Our conclusion compliments Libet's results (Libet et al., 1983, p. 632) which also showed the same trend. The most probable reason why the results do not replicate the order effect found in Dominik et al. (2017) is that in this case, the participants were already familiar with both the M and W tasks when making the W reports which were analyzed.

Libet's grand averaged S report was −47 ms. This is by 193 ms earlier than our mean S(A) reports and by 81 ms earlier than our mean S(O) report. In this case, we are prone to believe that the difference is indeed caused by omission of the S precision training.

However, if this is the case, the training seems to influence the S(A) reports far more than the S(O) reports. In fact, we believe that the O mode of recall might generally be more precise, given the fact that the (O) reports have universally lower SEM and that the mean S (O) and M(O) reports are notably closer to the target event (stimulus or the EMG onset) than the S(A) and M(A) counterparts. This also rules out the possibility that the strikingly late S(A) reports are caused by the unknown stimulator latency (see Section 4.1), because if this was the case, then the S(O) reports should have also been affected (there is no reason to assume that the stimulator would exhibit different latencies in the S(A) trials than in the S(O) trials). Nevertheless, the skin stimulus is delivered by moving the bolt towards the participant's skin and then retracting it – we can speculate that the S(A) reports could somehow lead the participant to assess the time of the "end of the touch", while the S(O) task would encourage the participant to report the time of the "beginning of the touch". This may require further testing using various skin stimulators.

We also found significant differences between the M(A), W(A) and S(A) reports. We find it important that the W(A) reports precede the M(A) reports significantly, suggesting that our participants did not usually confuse the W impression for the M impression. However, it should be noted that the M and W reports of participant 3 did not significantly differ, complimenting the idea presented in Dominik et al. (2017) suggesting that the W experience might have, for some individuals, the same meaning as the M experience.

By the end of this section, we feel obliged to point out one specific problem, which seems to significantly encumber the "libetian" discussion – the SEMs can be assessed using two different calculation methods which may lead to vastly different results. If the SEM is calculated as the standard deviation of the series means, then the procedure corresponds to Libet's original procedure. On the other hand, if the SEM is calculated using the formula:

$$SEM = \frac{SD}{\sqrt{n}}$$

then the procedure assumes that there is no difference between the series means (which is not true, as suggested in Table 6) and may lead to an underestimated result. We therefore recommend calculating the SEM using the first procedure when reporting the introspective data in Libet-style experiments.

4.5. ERP discussions

When analyzing the EEG data, we found several loopholes in Libet's original process. Many of the common RP analysis problems are discussed elsewhere (e.g. Verbaarschot et al., 2015). We will therefore adress only the problems which seem to be relevant to Libet's suggestions (Libet et al., 1982, 1983).

First and foremost, in many cases we did not find any RPs in the EEG plots where they were actually supposed to be found. Our experience can be summarized by a quotation of Pockett and Purdy (2011, pp. 4–5, our bold added): "*When one first begins to investigate the event-related potentials arising in the 2 s prior to voluntary movements, it rapidly becomes clear that* ***not all experimental subjects generate RPs****. As with many negative findings, the idea of trying to publish this result is soon overtaken by the realization that it would be far too easily rejected on the grounds that everyone can record RPs, so there must have been some technical inadequacy in the recording sessions where none was seen."* We are indeed aware of some technical limitations to our EEG recording equipment (see Section 5), but we would like to support Pockett and Purdy in their opinion and to suggest that in many circumstances, the averaged EEG plots might, indeed, not exhibit any RP detectable by an eye-ball inspection when it is in fact expected. When inspecting the EEG plots using the MN method (i.e. eye-ball inspection), we agreed to assess an RP only if it satysfyingly fitted the description in our "Guide to the RP identification" (see the supplement). Our reason for this was to be sure that we analyze an RP and not some non-specific negativities present even in a resting EEG (in fact, we found some RPs preceding the skin stimulation in the S series which makes it clear that the RP may be falsely identified where it is not supposed to be identified, even while doing it with a great caution). We used independent assessments made by five researchers, merged the assessments and then let another examiner check the results. However, the results might have been influenced by the fact that when merging, some assessments had to be corrected because they violated an important rule in the "Guide" or because the examiners did not agree on the onset timing. In both cases, the merging examiner had to subjectively choose one of the assessments, which admittedly introduced some amount of the previous subjectivity into the process.

Another potential factor reducing the number of identified RPs was relatively high high-pass filter (0.5 Hz), which was, however, found to be necessary to remove slow voltage drifts present in the raw recording, but could also move the onset notably (see Verbaarschot et al., 2015, Fig. 6). One could also wonder why we made the MN assessments with millisecond precision and not by 100 ms step as in the Libet's case (see Libet et al., 1983, p. 630, Table 1); the reason was that RPs sometimes had sharp onsets, so we found it useful to time the onsets precisely (which we did using a cursor which displayed the timestamp of a specific data point). To summarize the discussion of MN method, we are aware that some negativities might have in fact been RPs, but as it was often difficult to recognize an RP in an ambiguous plot, we chose to use a rather cautious approach.

We argue that the $RP_{90\%}$ method seems to be also problematic, due to multiple factors. First, it seems to detect the RP onset systematically later than it is supposed to be detected using the MN method (see Verbaarschot et al., 2015, Fig. 4A and B). Second, its result is heavily dependent on the baseline, which needs to be set prior to the calculation; we had to use different baselines for the M, W and S series than for the P and Pv series, because too early-set baseline in the W, M and S recordings distorted the onset estimates both ways due to early voltage shifts while the late baseline in the P and Pv series was calculated from a segment in which the RP negativity was already rising. We found that the $RP_{90\%}$ was more sensitive to negativities which were not identified as RPs in the MN method, but we find it questionable which of the two methods represent the reality better.

Even though the RPs were not found by the MN method in all relevant series, our recordings were indisputably valid, as argued in Section 3.3 and demonstrated in grand-averaged plots in Fig. 10. Most strikingly, there seems to a pronounced RP negativity in the M/W, P and Pv series while no such negativity is present in the grand-averaged EEG plot from the S series. It is also interesting to confirm Libet's assertion than an RP precedes a target time in the Pv series, even though no supraliminal movement was made. The grand-averaged RP in the Pv series (see Fig. 10.4) corresponds strikingly to Libet's description of the vetoed RP: "*In these series a ramplike [sic] pre-event potential was still recorded during greater than 1* sec *before the preset time (…), even though no actual muscle activation occurred (…). This resembles the RP of self–initiated acts when preplanning is present.*" (Libet, 1985, p. 538).

Before we analyzed the individual EEG plots, we defined the four RP types (see Section 2.6). However, we encountered a problem that Libet characterized the types by both their shape and their onset (Libet et al., 1982, p. 326), which led to many paradoxical findings of inconsistent shape and onset (e.g. a dome-shaped type II RP with an onset −300 ms). These inconsistencies were solved by prioritizing the shape to the onset, with an exception of type III RPs, which we defined as any late-onset negativities that cannot be regarded type II RPs. This might have lead to a rather large number of type III RPs found in our data compared to their scarcity in Libet's findings (Libet et al., 1983, p. 630, Table 1). Overall, we understand the benefit of classifying the RPs, but we did not find a clear connection between type I RP and the pre-planning impression or between type III RP and the impression of surprise by own movement (as suggested by Libet et al., 1982, pp. 327–330), even though it seems suggestive that the "surprise" impression may be connected to late RP onset and low amplitude (see Section 3.5). Our method of asking the participants whether they felt any pre-planning or spontaneity was rather crude; it might be advisable to use real-time inquiring as proposed by Verbaarschot et al. (2015, p. 304).

Our findings show that if an RP was found, it almost universally tends to precede the W report. However, Table 6 also shows that in some series of one participant, the mean W report was remarkably close to RP_{MN} onset (participant 2, sessions 3 and 5). It seems suggestive that the "*RP does not precede the intention to act in all participants*" (Verbaarschot et al., 2015, p. 310).

4.6. Libet's experiment interpretations discussions

Because our data differ notably from Libet's original outcomes, it seems advisable to discuss the potential implications for the validity of Libet's experiment. First and foremost, it seems to remain true that the RP generally precedes both the M and W reports acquired by the rotating-spot method. Libet et al. (1982; 1983) interpret the RP as a neural precursor of a decision to move. However, our outcomes do not rule out some alternative interpretations of the RP. For instance, Alexander et al. (2016) suggest that a negative potential remarkably similar to the RP can also be observed in a case of purely cognitive decision, which lacks any detectable muscle movement. Schurger et al. (2012) offerred an interpretation of the RP as a result of averaging spontaneous fluctuations in EEG occasionally building up to an activation threshold. However, as Schurger et al. point out, this interpretation pertains only RPs present in the task with self-initiated movement – thus, it cannot explain the RP waveform in the P and Pv series; nevertheless, this does not mean that spontaneous fluctuation cannot at least partially contribute to the RP generation in the self-inititated movement tasks. Because our results do not contradict these alternative explanations, the challenges for the Libet's interpretations implied by them remain valid and should be further investigated.

Another point of interest in this discussion is our confirmation of Libet's suggestion that the RP precedes a vetoed movement (Libet, 1985). This finding compliments conclusions by Schultze-Kraft et al. (2016), who also demonstrated that participants are able to veto a movement after the RP onset. In fact, their methodology seems to be even more convincing than Libet's original veto procedure, because the task by Schultze-Kraft et al. allowed the participants to prepare to move at any time and then ordered them to stop the movement when a BCI detected an RP onset (thus avoiding the need to make the movement "pre-set"). Libet (1985) suggests that the conscious veto can actually be considered an instrument allowing us to exert free will, because it can provide a way to block unconsciously arisen urge to act and because it presumably does not have any known neural correlate. However, complete freedom of such veto is in question, as it was demonstrated elsewhere that the conscious veto seems to have a neural correlate as well (Brass & Haggard, 2007). Therefore, even though we managed to replicate Libet's results in the Pv series, we do not have any evidence to support the claim that the conscious veto represents free will (or rather "free won't").

As stated in Section 1, the use of the rotating-spot method as means of measuring the introspective reports has been investigated in several papers. While the validity of the rotating-spot method for the M reports seems to be convincingly supported (Pockett & Miller, 2007), many studies suggested or demonstrated possible biases in the W reports, i.e. the reports of the urge, intention or wanting to move (e.g. Gomes, 1998; Lau, Rogers, & Passingham, 2007; Danquah et al., 2008; Banks & Isham, 2009). It is even possible that the W impression is not directly accessible to the consciousness and is a mere guess based on what the participant assumes the experimenter expects (see Dominik et al., 2017). Therefore, it is dubious at least to base any large claims about the free will on the compariosn of the W reports and RP onsets.

5. Limitations

Present study contains several technical limitations, some of which were already mentioned in the previous sections. Here, we will report additional important technical difficulties which may limit our study.

One limitation is a minor flaw in the mean report calculation from the M and W series in the order (O) mode of recall. The flaw is that not all trials in the O mode of recall contained an EMG onset, even though the button was pressed. The estimate might therefore be slightly inaccurate, because we related the reports to the mean EMG onset calculated for the whole series, but we assume that the error was not systematical. This limitation can be overcome in the future research if the EMG, not the mouse click, controled the

Libet's clock software (as in Libet's original experiment).

Other limitations pertain the EEG recordings and analyses. For the recordings, we used the available BIOPAC EEG100C amplifiers designed for spontaneous electroencephalography, not the ERS100C amplifiers designed for event-related potentials recordings. Although we preprocessed the EEG signal as stated in Section 2.6, the EEG contained a large amount of noise and many of the recordings had to be rejected. This might also be the reason why in many of the averaged EEG plots we failed to identify an RP, which could potentially be present but was obscured by the noise. The noise, together with the presence of eye movement artifacts, led to rejecting many single trials from recordings of overall satisfying quality. We do recommend using specialized equipment and combine the EEG recordings with EOG to identify and filter out the artifact caused by eye movement. We also recommend following additional EEG processing steps, such as re-referencing the signal (see Verbaarschot et al., 2015), many of which were unavailable to us, because we did not record the EEG from the whole scalp.

6. Conclusions

Our study's goal was to replicate Libet's experiment and to point out some methodological problems obscured by the experiment's complexity. To our knowledge, a replication as complex as ours was never carried out, and even though we failed to replicate some elements of Libet's experiment accurately, we find it critical to present our approach and outcomes. We showed some technical issues in both Libet's methodology and methodology used by later Libet-style experiments. Our data also showed that the results are highly variable and that the discrepancies found between many Libet-style studies are still found even when the original methodology is followed closely.

7. Funding

This research was supported by Palacký University Olomouc, Czech Republic [IGA FF_2017_021].

Acknowledgment

The authors would like to thank Michal Kluka for providing language help and Miroslav Kuba and Jan Kremláček from Faculty of Medicine in Hradec Králové, Charles University for consultations pertaining the ERP analyses.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.concog.2018.07.004.

References

Alexander, P., Schlegel, A., Sinnott-Armstrong, W., Roskies, A. L., Wheatley, T., & Tse, P. U. (2016). Readiness potentials driven by non-motoric processes. *Consciousness and Cognition, 39*, 38–47. https://doi.org/10.1016/j.concog.2015.11.011.

Banks, W. P., & Isham, E. A. (2009). We infer rather than perceive the moment we decided to act. *Psychological Science, 20*(1), 387–391. https://doi.org/10.1111/j.1467-9280.2008.02254.x.

Brass, M., & Haggard, P. (2007). To do or not to do: The neural signature of self-control. *The Journal of Neuroscience, 27*(34), 9141–9145. https://doi.org/10.1523/jneurosci.0924-07.2007.

Breitmeyer, B. G. (2002). In support of Pockett's critique of Libet's studies of the time course of consciousness. *Consciousness and Cognition, 11*, 280–283. https://doi.org/10.1006/ccog.2002.0555.

Caspar, E. A., & Cleeremans, A. (2016). "Free will": Are we all equal? A dynamical perspective of the conscious intention to move. *Neuroscience of Consciousness, 2015*(1), 1–10. https://doi.org/10.1093/nc/niv009.

Danquah, A. N., Farrell, M. J., & O'Boyle, D. J. (2008). Biases in the subjective timing of perceptual events: Libet et al. (1983) revisited. *Consciousness and Cognition, 17*, 616–627. https://doi.org/10.1016/j.concog.2007.09.005.

Dennett, D. C., & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences, 15*(2), 183–247. https://doi.org/10.1017/S0140525X00068229.

Dominik, T., Dostál, D., Zielina, M., Šmahaj, J., Sedláčková, Z., & Procházka, R. (2017). Libet's experiment: Questioning the validity of measuring the urge to move. *Conscioueness and Cognition, 49*, 255–263. https://doi.org/10.1016/j.concog.2017.01.017.

Dominik, T., Dostál, D., Zielina, M., Šmahaj, J., Sedláčková, Z., & Procházka, R. (2018). EEG Data and Introspective Reports from the Libet's Experiment Replication. Data in Brief.

Edwards, M. J., Moretto, G., Schwingenschuh, P., Katschnig, P., Bhatiaa, K. P., & Haggard, P. (2011). Abnormal sense of intention preceding voluntary movement in patients with psychogenic tremor. *Neuropsychologia, 49*, 2971–2973. https://doi.org/10.1016/j.neuropsychologia.2011.05.021.

Gomes, G. (1998). The timing of conscious experience: A critical review and reinterpretation of Libet's research. *Consciousness and Cognition, 7*, 559–595. https://doi.org/10.1006/ccog.1998.0332.

Gomes, G. (2002). Problems in the timing of conscious experience. *Consciousness and Cognition, 11*, 191–197. https://doi.org/10.1006/ccog.2002.0550.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*, 128–133.

Haggard, P., & Libet, B. (2001). Conscious intention and brain activity. *Journal of Consciousness Studies, 8*(11), 47–63.

Keller, I., & Heckhausen, H. (1990). Readiness potentials preceding spontaneous motor acts - Voluntary vs involuntary control. *Electroencephalography and Clinical Neurophysiology, 76*, 351–361.

Klein, S. (2002). Libet's research on the timing of conscious intention to act: A commentary. *Consciousness and Cognition, 11*, 273–279. https://doi.org/10.1006/ccog.2002.0557.

Klemm, W. R. (2010). Free will debates: Simple experiments are not so simple. *Advances in Cognitive Psychology, 6*, 47–65. https://doi.org/10.2478/v10053-008-0076-2.

Lau, H. C., Rogers, R. D., Haggard, P., & Passingham, R. E. (2004). Attention to intention. *Science, 303*(5661), 1208–1210. https://doi.org/10.1126/science.1090973.

Lau, H. C., Rogers, R. D., & Passingham, R. E. (2006). On measuring the perceived onsets of spontaneous actions. *The Journal of Neuroscience, 26*(27), 7265–7271. https://doi.org/10.1523/jneurosci.1138-06.2006.

Lau, H. C., Rogers, R. D., & Passingham, R. E. (2007). Manipulating the experienced onset of intention after action execution. *Journal of Cognitive Neuroscience, 19*(1), 81–90. https://doi.org/10.1162/jocn.2007.19.1.81.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences, 8*, 529–566.

Libet, B. (2000). Time factors in conscious processes: Reply to Gilberto Gomes. *Consciousness and Cognition, 9*, 1–12. https://doi.org/10.1006/ccog.1999.0408.

Libet, B. (2004). *Mind time.* Cambridge, MA: Harvard University Press.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential); The unconscious initiation of a freely voluntary act. *Brain, 106*, 623–642.

Libet, B., Wright, E. W., & Gleason, C. A. (1982). Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts. *Electroencephalography and clinical Neurophysiology, 54*, 322–335.

Moretto, G., Schwingenschuh, P., Katschnig, P., Bhatia, K. P., & Haggard, P. (2011). Delayed experience of volition in Gilles de la Tourette syndrome. *Journal of Neurology, Neurosurgery & Psychiatry, 82*, 1324–1327. https://doi.org/10.1136/jnnp.2010.221143.

Papanicolaou, A. C. (2017). The myth of the neuroscience of will. *Psychology of Consciousness: Theory, Research, and Practice, 4*(3), 310–320. https://doi.org/10.1037/cns0000116.

Picton, T. W. (1992). The P300 wave of the human event-related potential. *Journal of Clinical Neurophysiology, 9*(4), 456–479.

Pirio Richardson, S., Matsuhashi, M., Voon, V., Peckham, E., Nahab, F., Mari, Z., et al. (2006). Timing of the sense of volition in patients with schizophrenia. *Clinical Neurophysiology, 117*, S49–S111. https://doi.org/10.1016/j.clinph.2006.06.120.

Pockett, S. (2002). On subjective back-referral and how long it takes to become conscious of a stimulus: A reinterpretation of Libet's data. *Consciousness and Cognition, 11*, 144–161. https://doi.org/10.1006/ccog.2002.0555.

Pockett, S., & Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition, 16*(2), 241–254. https://doi.org/10.1016/j.concog.2006.09.002.

Pockett, S., & Purdy, S. (2011). Are voluntary movements initiated preconsciously? The relationships between readiness potentials, urges and decisions. In In W. Sinnott-Armstrong, & L. Nadel (Eds.). *Conscious will and responsibility: A tribute to Benjamin Libet* (pp. 34–46). New York: Oxford University Press. https://doi.org/10.1093/acprof: oso/9780195381641.001.0001.

Schultze-Kraft, M., Birman, D., Rusconi, M., Allefeld, C., Görgen, K., Dähne, S., et al. (2016). The point of no return in vetoing self-initiated movements. *Proceedings of the National Academy of Sciences of the United States of America, 113*(4), 1080–1085. https://doi.org/10.1073/pnas.1513569112.

Schurger, A., Sitt, J. D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences of the United States of America, 109*(42), E2904–E2913. https://doi.org/10.1073/pnas.1210467109.

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., et al. (2004). Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience, 7*(1), 80–84. https://doi.org/10.1038/nn1160.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Science, 11*(5), 543–545. https://doi.org/10.1038/nn.2112.

Trevena, J. A., & Miller, J. (2002). Cortical movement preparation before and after a conscious decision to move. *Consciousness and Cognition, 11*, 162–190. https://doi.org/10.1006/ccog.2002.0548.

Verbaarschot, C., Farquhar, J., & Haselager, P. (2015). Lost in time... the search for intentions and readiness potentials. *Consciousness and Cognition, 33*, 300–315. https://doi.org/10.1016/j.concog.2015.01.011.