Can hypnotic suggestibility be measured online?

Palfi, B. [1,2a], Moga, G. [1], Lush, P. [2,3], Scott, R. B. [1,2], & Dienes, Z. [1,2]

[1]School of Psychology, University of Sussex, Brighton, UK

[2]Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK.

[3]School of Informatics, University of Sussex, Brighton, UK

[a]To whom correspondence should be addressed: Bence Palfi

E-mail: b.palfi@sussex.ac.uk

Word count: 6063

Abstract

Hypnosis and hypnotic suggestions are gradually gaining popularity within the consciousness community as established tools for the experimental manipulation of illusions of involuntariness, hallucinations and delusions. However, hypnosis is still far from being a widespread instrument; a crucial hindrance to taking it up is the amount of time needed to invest in identifying people high and low in responsiveness to suggestion. In this study, we introduced an online assessment of hypnotic response and estimated the extent to which the scores and psychometric properties of an online screening differ from an offline one. We propose that the online screening of hypnotic response is viable as it reduces the level of responsiveness only by a slight extent. The application of online screening may prompt researchers to run large-scale studies with more heterogeneous samples, which would help researchers to overcome some of the issues underlying the current replication crisis in psychology.

Hypnosis and hypnotic suggestions have been shown to be useful experimental tools to test theories of cognitive neuroscience (Oakley & Halligan, 2013; Raz, 2011), especially theories related to consciousness (Cardeña, 2014; Terhune, Cleeremans, Raz, & Lynn, 2017). For instance, hypnotic suggestions can evoke changes in the feeling of voluntariness (Weitzenhoffer, 1974, 1980) or even modify one`s sense of agency (Haggard, Cartledge, Dafydd, & Oakley, 2004; Lush et al., 2017; Polito, Barnier, & Woody, 2013). Responses to suggestions frequently involve alterations in perception, such as the experience of positive and negative hallucinations or delusions (Kihlstrom, 1985; Oakley & Halligan, 2009). Moreover, hypnotic suggestions can be employed to simulate some properties of neurological and psychiatric conditions in healthy subjects (Barnier & McConkey, 2003; Oakley, 2006). Finally, correlations between hypnotisability and measures employed by consciousness researchers (e.g., the rubber hand illusion; the vicarious pain questionnaire; mirror touch synaesthesia) have recently been found (Lush et al., 2018). These correlations suggest that measures common in the consciousness literature are driven by hypnotic suggestibility. There is therefore an increasing need for an expansion of hypnosis research. Unfortunately, the successful application of hypnotic suggestions demands plenty of resources, making it impractical for researchers to run large-scale hypnosis related studies. In order to conduct experiments involving hypnosis, researchers generally need to recruit from a specific subsample of people based on their tendency to respond to hypnotic suggestions. To achieve this, researchers run hypnosis screening sessions before recruitment, so that, for example, they can identify the participants at the lowest and highest end of the scale (low and highly hypnotisable people, respectively). High and low hypnotisability are usually defined as the top and bottom 10%-15% of screening scores (Barnier & McConkey, 2004; Anlló, Becchio & Sackur, 2017). Therefore, screening procedures are time-consuming; to identify a single highly suggestible participant for an experiment, one has to find, on average, ten people who are willing to undertake a screening that can last from 40 up to 90 minutes depending on the applied method.

The hypnosis screening procedure has moved through a long developmental process in which it has become more and more user friendly. Initially, the screening consisted of two steps, a preliminary group session applying the Harvard Group Scale of Hypnotic Susceptibility Form A (HGSHS:A; Shor & Orne, 1963) and an individual session using the Stanford Hypnotic Susceptibility Scale Form C (SHSS:C; Weitzenhoffer & Hilgard, 1962) conducted with only those scoring very high or low in the first session. The later development of a reliable group screening method, the Waterloo-Stanford Group Scale of Hypnotic Susceptibility (WSGC; Bowers, 1993), has drastically mitigated the time required for screening as it allows researcher to screen up to a dozen people in about 90 minutes (although it was originally intended to act as a second screen after an HGSH:A, a single screen with the WSGC is quite reliable enough to select subjects capable of later having compelling subjective responses to difficult suggestions, e.g. digit-colour synesthesia, Anderson, Seth, Dienes, & Ward, 2014, or compelling objective reductions in Stroop interference to alexia suggestions, e.g. Parris, Dienes, Bate, & Gothard, 2014). Recently, the Sussex Waterloo Scale of Hypnotizability (SWASH; Lush, Moga, McLatchie & Dienes, 2018) was introduced, which is a modified version of the WSGC. The SWASH includes new items to measure the subjective experiences of the participants (compare also the Carleton University Responsiveness to Suggestion Scale

[CURSS, Spanos, Radtke, Hodgins, Stam, & Bertrand, 1983], and the Creative Imagination Scale [CIS, Wilson & Barber, 1978]). The length of the procedure was reduced to 40 minutes and it can be run with larger groups than the WSGC (Lush et al., 2018). Moreover, the dream and age regression suggestions were not included in the SWASH. These highly personalized items of the WSGC can be risky by virtue of possibly triggering unpleasant memories or emotions (Cardeña & Terhune, 2009; Hilgard, 1974).

Nonetheless, the application of the least demanding methods (such as the SWASH, the CURSS or the CIS), still requires potential participants to attend a group session, which makes the screening procedure relatively time consuming and limits the subject pools to psychology students who are the easiest to incentivise to participate in a group screening on campuses. These two barriers of large-scale hypnosis studies could be overcome by employing fully automatized, online hypnosis screening procedures. In the last two decades, psychological science has witnessed growth in the application of online data collection for experimental purposes, paving the way for researchers to collect large samples in a short period of time (Reips, 2000; though it can come with its own problems, e.g. Dennis, Goodson & Pearson, 2018). In order to adapt the hypnosis screening procedure online, one needs to ensure that the non "live" version can induce similar objective and subjective hypnotic responses as with a "live" hypnotist. Indeed, suggestibility scores of participants are comparable when the hypnotic induction and suggestions are delivered by a pre-recorded audiotape and when they are delivered by an experimenter (Barber & Calverley, 1964; Fassler, Lynn, & Knox, 2008; Lush, Scott, Moga, & Dienes, 2018). These findings underpin the idea that the participants could easily undergo a hypnosis screening procedure in their own rooms by listening to a pre-recorded script and filling out the booklets online. Nevertheless, online data collection has its own perils, namely, the data acquired by online questionnaires might not be as reliable and the results might not be consistent with the ones of the traditional data collection procedures (Krantz & Dalal, 2000). Therefore, the reliability of new online questionnaires, such as the online version of a hypnosis screening procedure, needs to be tested even if there is evidence that the quality of the data and the findings of online based studies can be similar to those obtained by traditional methods (Gosling, Vazire, Srivastava, & John, 2004; Buhrmester, Kwang, & Gosling, 2011).

In this project, our purpose is to explore the extent to which an online hypnotic screening procedure is reliable and consistent with an offline procedure. To this aim, we measured people`s hypnotic suggestibility with the SWASH on two separate occasions and in two different environments. Henceforth, we call every type of data collection carried out in a controlled environment with the experimenter present an offline screening, whereas undertaking a hypnotic screening alone in one's own room under one's own control will be called online screening. In addition, we are interested in the extent to which the length of the delay between first and second screen can influence the reliability and the scores of hypnotic suggestibility. The question about the stability of hypnotic suggestibility over periods of few days or even decades have inspired various research projects (e.g., Fassler, Lynn, & Knox, 2008; Lynn, Weekes, Matyi, & Neufeld, 1988; Piccione, Hilgard, & Zimbardo, 1989). To assess the stability of hypnotic suggestibility, we recruited half of the sample from the subject

pool of the year of 2016 and the other half from the year of 2017, both of whom have already received offline screening. Therefore, for some of the participants, the delay between the two screenings is not more than 6 months (short delay group), whereas for the others, it is at least one and a half years (long delay group). For practical reasons, the first screening was organised offline, in groups of 20-40 for all the participants, whereas the second screening was either an online screening or another offline one. By this method, we are able to estimate how strongly the type of the screening and the length of the delay can influence the suggestibility scores of the people; we can also assess their influence on the test-retest reliability and the validity of the screening. Taken together, this project strives to explore whether a well-established offline screening procedure could be replaced for practical purposes by an online version, which could help consciousness researchers run more and larger hypnosis studies by drastically cutting the recruitment related costs.

While responding to hypnotic suggestions, people tend to experience as of being in some form of trance or altered state (Kihlstrom, 2005; Kirsch, 2011). This experience is usually measured by subjective reports of depth of hypnosis (e.g., Hilgard & Tart, 1966), which is, interestingly, strongly associated with people`s ability to respond to hypnotic suggestions (Wagstaff, Cole, & Brunas-Wagstaff, 2008). We investigate this link by assessing the strength of relationship between hypnotic suggestibility scores and depth of hypnosis reports, and the extent to which the mentioned experimental manipulations can influence this relationship. We also aim to evaluate the extent to which depth of hypnosis is influenced by the type of data collection and the length of the delay between screens to ensure that people experience comparable level of hypnotic depth during online and offline screens.

In our analyses, we solely employed estimation procedures instead of testing the existence of differences with an inferential statistical tool such as the null-hypothesis significance test (Fisher, 1925; Neyman & Pearson, 1933) or the Bayes factor (e.g., Dienes, 2011; Rouder et al., 2009). Estimation is recommended over inferential statistics when the existence of a difference is established or it is not relevant (Jeffreys 1961; Wagenmakers et al., 2018). The second point proves to be decisive for our case, since it is not necessary to test the existence of any investigated effect to answer our research questions. For instance, the core aim of the current project was to conclude regarding the applicability of online hypnosis screening by comparing the SWASH scores, the reliability and the validity of online and offline hypnosis screening. Imagine a scenario in which an inferential statistical tool demonstrates evidence for the difference between the offline and online groups in favour of the offline group in all aspects that assess the quality of the measurement. Importantly, this outcome per se cannot give a definite answer to our central question as the mere fact that offline screening is significantly better than online screening neglects the question of magnitude of the difference. To reject or accept the idea that online screening is viable, we need to know the extent to which the quality of offline and online screening differs so that we can decide whether the benefits of the online screen outbalance its costs. Further, the fact that the two types of screening will correlate cannot be in doubt; the question is simply the strength of the relationship between them.

To explore the range of plausible effect sizes, estimation methods, either from the Bayesian (Kruschke, 2010, 2013; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Wagenmakers, Morey, & Lee, 2016) or from the frequentist school (Cumming, 2014) can be used. Here, we applied a Bayesian tool, estimation by calculating the 95% Bayesian Credibility Intervals, as this is the method that is appropriate to answer our research question; namely, how confident can we be that the true effect size lies within a specific interval (Wagenmakers et al., 2018). Only Credibility Intervals allow us to make claims such as that the true value of the effect size is probably not larger or smaller than a particular value.

Methods

Participants

Psychology students at the University of Sussex participated in an offline hypnosis screening as part one of their modules during the first semester of their studies. We recruited psychology students who had started their BSc studies in the year of 2016 or 2017 and who had provided their contact information in an offline hypnosis screening session. Both subject pools consisted of around 300 students and we randomly assigned half of them to each experimental group (experimental groups described below). Thus, we invited around 150 people for each group. We continued data collection until the end of the spring semester of 2018. In the second session, 73 students participated. However, we could not trace back the data of two students to their first session results and so we needed to exclude them from all of the analyses, leaving us with 71 participants in total. Twenty-six students attended the offline session (23 females, M_{age} = 19.7, SD_{age} = 1.8) and 45 students completed the screening online (41 females, M_{age} = 21.0, SD_{age} = 5.3).

We informed each participant about the nature of the study and only those students were be able to attend who agreed to the terms and conditions of the study. After finishing the experiment, the participants were debriefed and received a payment of £5 or course credit. The study has been approved by the Ethical Committee of the University of Sussex (Sciences & Technology C-REC).

Materials

One of the authors produced the audio recording of the hypnosis procedure (induction and the suggestions); the length of this recording was 28 minutes. The questionnaire applied in the first session for data collection was created in MatLab (MathWorks, 2016), whereas the questionnaire that was used in the second session was a PHP based website. The PHP script, the materials and the documentation on how to install the software can be accessed at https://osf.io/6twdp/.

Measures

The measures introduced below were utilised in the first occasion of the data collection. The second occasion only included the assessment of the hypnotic suggestibility measured by the SWASH regardless of the type of the session (offline or online). Note that, although several

questionnaires were registered along with the first screening, we only used the suggestibility scores of the participants in this project (see our research questions in the last paragraph of the Introduction).

SWASH. The hypnotisability of the students was measured by the SWASH. This scale is a modified version of the WSGC (Bowers, 1993) which contains 10 suggestions and corresponding items measuring objective suggestibility and the subjective experiences of the participants about each suggestions.

Data collection in 2016. As part of the first session in 2016 the following four questionnaires were registered: (a) Barratt Impulsiveness Scale (BIS-11), which consists of 30 items and measures people`s tendency to behave impulsively (Patton & Stanford, 1995); (b) Free Will Inventory (FWI), which includes 29 items measuring people`s beliefs about free will and their relationships with these beliefs (Nadelhoffer, Shepard, Nahmias, Sripada, & Ross, 2014); (c) Short Form of the Five Facet Mindfulness Questionnaire (FFMQ-SF), which is a 24-item scale assessing the mindfulness skills of individuals via self-report (Bohlmeijer, ten Klooster, Fledderus, Veehof, & Baer, 2011); (d) Dissociative Experiences Scale-II (DES-II), which is a 28-item self-report questionnaire developed by Bernstein and Putman (1986).

Data collection in 2017. In 2017, we administered the following four questionnaires in the first data collection session of: (a) a 15 minutes long breath counting exercises based on Study 2 of Levinson, Stoll, Kindy, Merry & Davidson (2014); (b) the Mindful Attention Awareness Scale (MAAS) consisting 15 Likert scale items (Brown & Ryan, 2003); (c) the Schizotypal Personality Questionnaire-Brief (SPQ-B), which consists of dichotomous questions (Raine & Benishay, 1995); (d) the DES-II that was used in 2016.

Design

We employed a 2*2*2 mixed design. The within subject variable is the date of the data collection (first session vs. second session). The between subjects independent variables are the form of the second hypnosis screening session (offline vs. online) and the length of the delay between the first and the second sessions (short delay [few months] vs. long delay [more than a year]).

Procedure

There were three forms of data collection: 1) group sessions at the university with the experimenter present (first, offline screen) 2) individual sessions in a small experimental room at the university with the experimenter present (second, offline screen) 3) individual sessions at home (second, online screen,). All of the participants engaged in the first, offline, screen and later they were invited to attend in a second screen that was either offline or online. The procedure of the screening was identical in each case and followed the steps below.

After providing informed consent, the participants had the opportunity to provide contact details for a database in case they were willing to participate in hypnosis related research in the future. Next, they were asked to adjust the volume of their headphones until it

was moderately loud by listening to a test tune. Before starting the hypnotic induction procedure, they were notified that the whole procedure would last about 45 minutes and that they should not take a break. By pressing the start button, participants ran the hypnotic induction and suggestions. After the de-induction, participants were asked to fill out the SWASH response booklet, rating their response to each suggestion. Finally, the participants were thanked for attending and debriefed.

Data analysis

Data transformation. We computed the Objective and Subjective suggestibility scores of the participants as described in the SWASH manual (Lush et al., 2018) and then we doubled all subjective scores so that both of the objective and subjective scores fell between 0 and 10. By taking the weighted average of these derived scores, we calculated the composite SWASH score of each participant, which was used in the majority of the analyses. For more details on the calculation of the SWASH scores, see Lush et al. (2018). Given that the distributions of the objective, subjective and composite SWASH scores of the first screen were all fairly normal (see Figure 1 in Preregistration), we assumed that the dataset of the second screen was also normally distributed. Therefore, we planned to use parametric methods to estimate the strength of correlation between continuous variables (Pearson`s *r*).

Bayesian estimation. In this project, we estimated the population effect sizes and did not test hypotheses. Thus, here, we report the estimates (e.g., mean or correlation) and the 95% Bayesian credibility intervals (CI) applying a uniform prior distribution. Note that, although, the bounds of the CIs are numerically equal to the bounds of the confidence intervals (assuming a uniform prior), their interpretation is different (e.g., Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).

Implementation of the preregistration

The design and research questions of this study were preregistered at osf.io/3abje. In order to ensure the reproducibility of the analysis and decrease analytic flexibility, we preregistered an analysis script, written in R (R Core Team, 2016), *a prior* to data collection. The script includes all of the steps defined in the preregistration and an additional data simulation, which helped us test and debug the script. In this paper, we present the results of analyses that were preregistered in the above-mentioned R script and results of two additional, non-preregistered analyses: 1) test-retest reliability of SWASH scores; 2) correlation between SWASH and depth of hypnosis scores. We deviated from the analysis script in one aspect. The calculation of the 95% CIs of the differences between two correlations was incorrect in the original script due to an issue with back-transformation of Fisher`s *z* values of difference scores to Pearson`s *r* (e.g., Meng, Rosenthal & Rubin, 1992; Olkin & Finn, 1995). Therefore, we used the cocor R package (Diedenhofen & Musch, 2015), which is based on the approximation method of Zou (2007), to estimate the 95% CIs of the differences between correlations.

Results

SWASH scores

The mean of the composite SWASH scores in the offline group (M = 3.44) was only slightly larger than the mean of the online group (M = 3.13) rendering their difference negligible (M_{diff} = 0.31, 95% CI [-0.59, 1.22]). Crucially, the difference between the groups is unlikely to be larger than 1.22. The difference between the offline and online groups is likely to be negligible or small for both of the objective (M_{diff} = 0.39, 95% CI [-0.58, 1.36]) and subjective subscales (M_{diff} = 0.24, 95% CI [-0.72, 1.19]). Panel A of Figure 1 demonstrates that the distribution of the composite SWASH scores of the offline group is akin to the online group. The density of the data is similar between the groups even around the right tail (top) of the distribution indicating that similar proportion of the participants scored high on the SWASH in the offline and online groups. The mean of the composite SWASH scores was comparable in the short (M = 3.32) and long delay groups (M = 3.10), and the plausible values of their differences vary around zero with a maximum difference of 1.10 (M_{diff} = 0.21, 95% CI [-0.67, 1.10]). Table 1 presents the means and 95% CIs of all groups and comparisons with the composite, objective and subjective scores separately.

Table 1.

The Mean Composite, Objective and Subjective SWASH Scores with 95% CIs Broken Down by the Type of the Second Screen and the Length of the Delay

Group	Measure		
	Composite	Objective	Subjective
Offline	3.44 [2.73, 4.16]	3.92 [3.17, 4.67]	2.96 [2.19, 3.73]
Online	3.13 [2.54, 3.71]	3.53 [2.89, 4.17]	2.72 [2.13, 3.32]
Difference	0.31 [-0.59, 1.22]	0.39 [-0.58, 1.36]	0.24 [-0.72, 1.19]
Short delay	3.32 [2.72, 3.91]	3.89 [3.26, 4.52]	2.74 [2.14, 3.35]
Long delay	3.10 [2.42, 3.78]	3.28 [2.54, 4.02]	2.93 [2.19, 3.67]
Difference	0.21 [-0.67, 1.10]	0.61 [-0.34, 1.57]	-0.19 [-1.12, 0.75]

Note. Values within the squared brackets represent the 95% Confidence Intervals. Data presented in this table are based solely on the second screen.

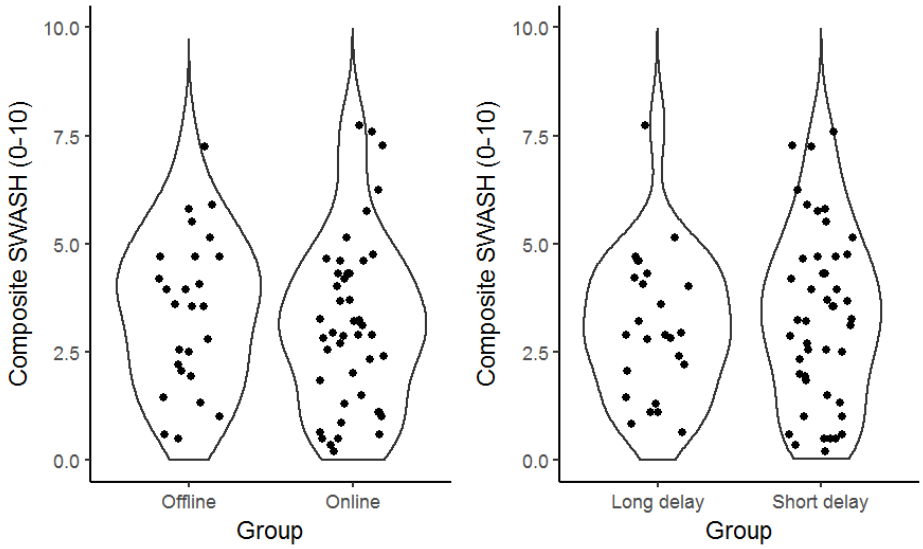

Figure 1. Violin plots depicting the distribution of composite SWASH scores of the second screens broken down either by the type of the screen (offline vs. online, panel A) or by the length of the delay (long vs. short delay, panel B). Each black dot indicates a composite SWASH score of a participant.

Validity

The correlation between the objective and subjective subscales of the SWASH was strong for the offline screen (r = .78, 95% CI [.56, .89]) as well as for the online screen (r = .79, 95% CI [.65, .88]) indicating appropriate validity in this respect. The difference between the offline and online screen in terms of the strength of the correlation between the objective and subjective scales was close to zero (r = - .02, 95% CI [-.25, .17]).

Test-retest reliability (non-preregistered)

Correlation between the first and the second screen scores was strong for the subjective subscale but only moderate for the objective subscale irrespective of the type of the screen. For the composite scores, the correlation was strong for the online and offline group as well indicating a good enough test-retest reliability of the SWASH. Interestingly, the test-retest reliability of the online group was possibly higher than that of the offline group, although, only to a small extent (See Table 2 for *r*s and their 95% CIs). The correlation between the first and second screen scores was strong in the short delay group for the subscales as well as for the composite scores. However, the correlation was only moderate in the long delay group implying that the test-retest reliability of the SWASH is influenced by the length of the delay between the screens from a weak to a moderate extent. Table 2 presents the exact correlation values and their 95% CIs separately for the experimental groups.

Table 2.

Test-retest Reliability of SWASH Scores Broken Down by Type of Screen and Length of Delay

Group	Measure		
	Composite	Objective	Subjective
Offline	.62 [.31, .81]	.43 [.05, .70]	.69 [.42, .85]
Online	.74 [.57, .85]	.59 [.35, .75]	.77 [.61, .87]
Difference	-.12 [-.45, .14]	-.16 [-.57, .20]	-.07 [-.37, .15]
Short delay	.79 [.65, .88]	.65 [.44, .79]	.81 [.68, .89]
Long delay	.55 [.20, .78]	.37 [-.03, .67]	.56 [.21, .78]
Difference	.24 [-.02, .61]	.28 [-.08, .70]	.25 [-.01, .61]

Note. The correlation values are all Pearson`s *r*s and the 95% CIs are reported within the squared brackets.

Depth of hypnosis

Difference between the groups. The participants reported somewhat higher depth of hypnosis scores in the offline (*M* = 2.15, 95% CI [1.61, 2.70]) than in the online (*M* = 1.73, 95% CI [1.31, 2.16]) group. Nonetheless, the difference between the groups is not substantial and the maximum plausible value of this difference is 1.10 (*M* = 0.42, 95% CI [-0.26, 1.10]). The mean of the depth of hypnosis scores in the short delay (*M* = 1.80, 95% CI [1.35 – 2.26]) compared to the long delay group (*M* = 2.04, 95% CI [1.59, 2.49]) differed only to a small extent (*M* = -0.24, 95% CI [-0.87, 0.40]). Figure 2 portrays the distribution of the depth of hypnosis scores broken down by the type of the second screen (panel A) and the length of the delay between the first and second screen (panel B). The depth of hypnosis scores are similarly distributed in the offline and online groups.

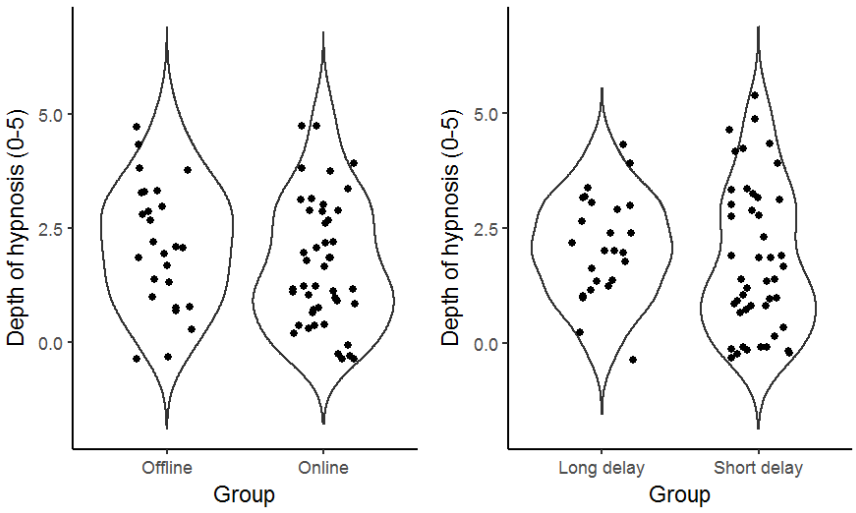

Figure 2. Violin plots representing the distribution of depth of hypnosis scores separately for the offline and online screens (panel A), and for the short and long delay groups (panel B).

Correlation between SWASH and depth of hypnosis scores (non-preregistered). The correlation between the SWASH and depth of hypnosis scores was strong for all but one measure in the online and for all in the offline screen group (all $r > .54$). The strength of the correlation is unlikely to be larger than .21 in the offline group than in the online group rendering the difference between the two groups minimal. There was strong correlation between the depth of hypnosis scores and all measures in the short delay group (all $r > .70$), and the correlations were moderate to strong in the long delay group (all $r > .31$). The difference between the two groups for the strength of the correlations was weak to moderate, and it was the highest for the objective scores. Table 3 shows all of the correlation values and their 95% CIs separately for the experimental groups and for all of the measures.

Table 3.

Correlation Between SWASH and Depth of Hypnosis Scores Broken Down by the Type of Screen and the Length of Delay

	Measure		
Group	Composite	Objective	Subjective
Offline	.76 [.53, .89]	.66 [.37, .84]	.77 [.54, .89]
Online	.70 [.52, .83]	.54 [.29, .72]	.81 [.67, .89]
Difference	.06 [-.21, .28]	.12 [-.22, .43]	-.04 [-.28, .15]
Short delay	.79 [.65, .88]	.70 [.51, .82]	.83 [.71, .90]
Long delay	.55 [.20, .78]	.31 [-.10, .63]	.70 [.41, .86]
Difference	.24 [-.03, .60]	.38 [.02, .81]	.13 [-.07, .42]

Note. The correlation values are all Pearson`s *r*s and the 95% CIs are reported within the squared brackets.

Discussion

The purpose of the present study was to explore whether online hypnosis screening is feasible as the adaptation of this method could ease the recruitment related costs of hypnosis research. To this aim, we estimated the extent to which offline and online hypnosis screening scores, measured by the SWASH, are comparable. The results revealed that the difference between offline and online groups was small to negligible in all aspects and, importantly, applying online rather than offline screening is unlikely to reduce the composite screening score by more than 1.22 and the objective score by more than 1.36 out of ten. To put these effect sizes in perspective, for instance, a recent meta-analysis of four studies investigating the influence of standard induction procedures on suggestibility found that, on average, people score 1.46 higher (out of ten) on scales assessing objective responses to suggestions if they had received

a priori induction compared to no induction (Martin & Dienes, 2018). Moreover, the average SWASH score in the online group was comparable to the result of an earlier screen conducted in group sessions at the same university (Lush et al., 2018). Finally, it is not only the average scores in the online group that can be deemed acceptable, the distribution of SWASH scores were also akin in the offline and online groups even at the positive end of the scale. This implies that some people can successfully respond to many suggestions when they undertake an online screening (See Figure 1). None of this was obvious before the data were collected.

The correlation between objective and subjective scores was strong for both of the offline and online groups; crucially, the correlation in the online group can only be as small as .65. This indicates that the validity of the SWASH remained acceptable even with online data collection. Moreover, the strength of the correlation between the subjective and objective components of the SWASH found by Lush et al (2018) was .70, which is consistent with our results. The strength of the correlation between SWASH scores of the first and second screens was medium in the offline and strong in the online group. The lower bound of the 95% CI in the online group was .57 implying that the test re-test reliability of the online measurement is adequate. These values are also appropriate in relative terms. For instance, Fassler et al. (2008) employed the CURSS which has an objective and a subjective subscale such as the SWASH, in two occasions and the test re-test correlations were .59 and .77 for the objective and subjective components, respectively. These results are in line with the correlations found by us in the online group. Overall, the psychometric properties of online screening were excellent; the quality of data collected online has shown to be consistent with the quality of offline data gathered within this study and as part of earlier studies with the SWASH and other hypnosis screening tools.

Modern theories of hypnosis advocate the notion that all hypnosis is self-hypnosis, since the hypnotic subject is the one who actively responds to the suggestions and creates the requested experience (Kihlstrom, 2008; Raz, 2011). This does not mean, however, that the experimenter has no influence on the responsiveness of the subject. For instance, the presence of an experimenter can be helpful in building up a rapport and facilitating responsiveness of the participants (e.g., Gfeller, Lynn, & Pribble, 1987). Nonetheless, the experimenter can also bias the responses of the subjects (e.g., Barber & Calverley, 1966; Troffer & Tart, 1966), and importantly, this level of bias can strongly vary across participants as it is almost impossible to deliver the induction and suggestions in an identical way multiple times. Therefore, the application of fully automatized screenings, such as the online version, can subserve the standardization of the assessment of hypnotic suggestibility.

Introducing online hypnosis screening would markedly decrease the amount of time experimenters need to invest to find participants for their studies. However, to complete a screening procedure, the participants still need to spend 45-60 minutes without taking a break; otherwise, the data would be not usable for recruitment purposes. A substantial part of the screening is assigned to the standard hypnotic induction, which consists of various suggestions mostly to relax; however, the responses to these suggestions are not assessed directly during the screening (e.g., Shor & Orne, 1963; Weitzenhoffer & Hilgard, 1962). Would it be feasible to exclude the standard induction from the screening procedure to save time for the

participants? Cognitive theories of hypnosis, such as the cold control theory (Barnier, Dienes, & Mitchell, 2008; Dienes & Perner, 2007), emphasise the role of the feeling of involuntariness in differentiating hypnotic from non-hypnotic responses. This feeling is also known as the "classical suggestion effect" (Weitzenhoffer, 1974, 1980). Therefore, according to cold control theory, not the practice of induction, but the feeling of involuntariness is the demarcation criterion, and it is important to ensure with self-report measures that the participants experienced a reduction in the level of control over their own behaviour (e.g., Palfi, Parris, McLatchie, Kekecs, & Dienes, 2018)[1]. From a practical perspective, it is important to bear in mind that the presence of a standard induction can increase responsiveness to the suggestions in the screening, in average, by 1.46 (Martin & Dienes, 2018) compared to the absence of the induction; and that the strength of the effect of an induction fluctuates across suggestions (Terhune & Cardeña, 2016). Nonetheless, as argued earlier in this paper, a general reduction of responsiveness does not qualify as decisive argument for retaining the induction procedure. As long as the absence of the induction does not produce a floor-effect or alters markedly the ranking of the suggestibility scores, the screening can be perfectly adequate for screening people for individual differences in response. Indeed, there are existing attempts to assess responsiveness to suggestions without exposing the participants to an induction, such as the Barber Suggestibility Scale (Barber & Glass, 1962) and the CIS (Wilson & Barber, 1978). These scales can be easily administered in a context presented as a test of imagination while applying motivational instructions to replace the induction or simply leaving out the induction. The existing evidence suggest that employing motivational instructions creates similar level of responsiveness as the application of the induction; however, the absence of the induction significantly dwindles the level of responsiveness to suggestions (Barber & Wilson, 1978). Future research could explore the extent to which the exclusion or replacement of the induction from the SWASH would be feasible and assess whether it would be beneficial.

A secondary interest of the current study was to assess the extent to which the length of the delay between the first and second screening affects the outcome of the screen and the psychometric properties of the measurement tool. Repeated assessment of suggestibility can negatively affect the suggestibility scores, for instance, if the delay amid the two occasions takes only a few days or weeks (Barber & Calverley, 1966; Fassler et al., 2008; Lynn et al., 1988). This reduction in suggestibility may be caused by boredom; the participants can become disengaged with the procedure by virtue of finding it repetitive (Barber & Calverley, 1966; Fassler et al., 2008). In our case, the short delay was a minimum of 5 months and we found no indication of substantial differences between the short and long delay groups among the SWASH subscales. For instance, Fassler et al. (2008) found a difference of 0.77 on the objective scores between the first and second session[2], but according to our data, the largest plausible difference is only .34. Nonetheless, the effect of boredom on the subjective scores

[1] An operational definition of hypnosis necessitates to usage of induction to render suggestions hypnotic, and labels all suggestions without a priori induction imaginative suggestion (Braffman & Kirsch, 1999; Kirsch, 1997; Kirsch & Braffman, 2001). This line of thinking would preclude us from omitting the induction in case we want to measure hypnotic suggestibility.

[2] The reported raw difference was 0.54; however, we adjusted this value from a scale of 0-7 to the scale of the SWASH, which is 0-10.

observed by Fassler et al. (2018) was 1.05[3], which is compatible with our results as the lower bound of the difference in that aspect was 1.12. Taken together, our data imply that the negative effect of boredom might wear off or becomes negligible after 5 months; however, more research is needed to settle this matter and identify the ideal amount of delay that can prevent boredom effects in repeated designs.

We note that our sample was restricted to university students, which might preclude the generalization of our findings, crucially, the applicability of online hypnosis screening, to a wider population. Nonetheless, the problem of generalizability represents a universal issue in experimental hypnosis research. For instance, a meta-analysis on 27 studies investigating hypnotically induced analgesia found that from the studies with non-clinical samples (N = 19), only one was run with people recruited from the local community whereas all the other studies were run with students (Montgomery, Duhamel, & Redd, 2000). Recruiting from a wider population would not only increase generalisability of the findings but it would further facilitate researchers to run large-scale hypnosis studies strengthening the replicability of the findings. Future research is needed to explore the extent to which online hypnosis research can be applied to screen and recruit people from local communities.

Finally, the vast majority of our participants were females; hence, the gender imbalance in our sample might be another factor hindering the generalizability of our findings. Research on the link between gender and hypnotic suggestibility has provided ambiguous results with some studies finding virtually no effect (Cooper & London, 1966; Dienes, Brown, Hutton, Kirsch, Mazzoni, & Wright, 2009; McConkey, Barnier, Maccallum, & Bishop, 1996) and some studies demonstrating a small effect size (Green, 2004; Green & Lynn, 2010; Morgan & Hilgard, 1973; Page & Green, 2007; Rudski, Marra, & Graham, 2004). Studies showing a small effect size of gender consistently found that women score higher than men, which might be a consequence of divergence in a personality trait that partly underlies suggestibility or difference between women and men in how they assess the difficulty of the suggestions (Rudski, Marra, & Graham, 2004). Nonetheless, these explanations are conjectures that have yet to be tested. With only seven men in the current data set, we can only speculate how much gender might moderate the difference between the online compared to the offline measurement of hypnotic suggestibility.

Conclusion

Altogether, the online assessment of hypnotic suggestibility appears to be feasible and the benefits far outweigh the downsides involved with its application. Although, online screening might be less engaging than the traditional, offline measurement of suggestibility and so it can result in slightly lower suggestibility scores, our study suggests that the effect size of this negative impact lies within acceptable boundaries. Crucially, the application of online hypnosis screening can subserve the execution of large-scale data collection with heterogeneous samples consisting of student and non-student participants as well. Furthering our knowledge based on small sample studies comes with many risks (e.g. Loken & Gelman, 2017), but the relative

[3] The reported raw difference was 2.2; we adjusted this value as well from a scale of 0-21 to the scale of the SWASH in which values can vary between 0 and 10.

high cost of hypnosis screening procedures hinders the researchers of the field from running well-powered studies. Therefore, we argue that the adaptation of online hypnosis screening is salutary and it helps experimental hypnosis research to realise its full potentials.

Acknowledgements

The preregistration, the data and the analysis script can be retrieved from https://osf.io/c46xa/. The authors declare no financial conflict of interest with the reported research. The project was not supported by any grant or financial funding. Bence Palfi is grateful to the Dr Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science.

References

Anderson, H. P., Seth, A. K., Dienes, Z., & Ward, J. (2014). Can grapheme-color synesthesia be induced by hypnosis?. *Frontiers in human neuroscience*, *8*, 220.

Anlló, H., Becchio, J., & Sackur, J. (2017). French norms for the Harvard Group Scale of hypnotic susceptibility, form A. *International Journal of Clinical and Experimental Hypnosis*, *65*(2), 241-255.

Barber, T. X., & Calverley, D. S. (1964). Comparative effects on" hypnotic-like" suggestibility of recorded and spoken suggestions. *Journal of Consulting Psychology*, *28*(4), 384.

Barber, T. X., & Calverley, D. S. (1966). Toward a theory of hypnotic behavior: experimental evaluation of Hull's postulate that hypnotic susceptibility is a habit phenomenon 1. *Journal of Personality*, *34*(3), 416-433.

Barber, T. X., & Glass, L. B. (1962). Significant factors in hypnotic behavior. *The Journal of Abnormal and Social Psychology*, *64*(3), 222.

Barber, T. X., & Wilson, S. C. (1978). The Barber suggestibility scale and the creative imagination scale: Experimental and clinical applications. *American Journal of Clinical Hypnosis*, *21*(2-3), 84-108.

Barnier, A. J., Dienes, Z., & Mitchell, C. J. (2008). How hypnosis happens: New cognitive theories of hypnotic responding. In M. Heap., R. J. Brown & D. A. Oakley (Eds.), *The Oxford handbook of hypnosis: Theory, research, and practice* (pp. 141-177). London: Routledge.

Barnier, A. J., & McConkey, K. M. (2003). Hypnosis, human nature, and complexity: Integrating neuroscience approaches into hypnosis research. *International Journal of Clinical and Experimental Hypnosis*, *51*(3), 282-308.

Barnier, A. J., & McConkey, K. M. (2004). Defining and identifying the highly hypnotizable person. *The highly hypnotizable person: Theoretical, experimental and clinical issues*, 30-61.

Bernstein, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *The Journal of Nervous and Mental Disease*, *174*(12), 727–735.

Bohlmeijer, E., ten Klooster, P. M., Fledderus, M., Veehof, M., & Baer, R. (2011). Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment, 18*(3), 308–320.

Bowers, K. S. (1993). The Waterloo-Stanford Group C (WSGC) scale of hypnotic susceptibility: Normative and comparative data. *International Journal of Clinical and Experimental Hypnosis, 41*(1), 35–46.

Braffman, W., & Kirsch, I. (1999). Imaginative suggestibility and hypnotizability: an empirical analysis. *Journal of Personality and Social Psychology*, *77*(3), 578.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology*, *84*(4), 822.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Cardeña, E. (2014). Hypnos and psyche: How hypnosis has contributed to the study of consciousness. *Psychology of Consciousness: Theory, Research, and Practice*, *1*(2), 123.

Cardeña, E., & Terhune, D. B. (2009). A note of caution on the Waterloo-Stanford Group Scale of Hypnotic Susceptibility: A brief communication. *Intl. Journal of Clinical and Experimental Hypnosis*, *57*(2), 222-226.

Cooper, L. M., & London, P. (1966). Sex and hypnotic susceptibility in children. *International Journal of Clinical and Experimental Hypnosis*, *14*(1), 55-60.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). Mturk Workers' Use of Low-Cost "Virtual Private Servers" to Circumvent Screening Methods: A Research Note.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on?. *Perspectives on Psychological Science*, *6*(3), 274-290.

Dienes, Z., Brown, E., Hutton, S., Kirsch, I., Mazzoni, G., & Wright, D. B. (2009). Hypnotic suggestibility, cognitive inhibition, and dissociation. *Consciousness and cognition*, *18*(4), 837-847.

Dienes, Z., & Perner, J. (2007). Executive control without conscious awareness: the cold control theory of hypnosis. In Jamieson, G. (Ed.), *Hypnosis and conscious states: The cognitive neuroscience perspective*, (pp. 293-314). Oxford University Press.

Fassler, O., Lynn, S. J., & Knox, J. (2008). Is hypnotic suggestibility a stable trait? *Consciousness and Cognition*, *17*(1), 240–253.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, Scotland: Oliver & Boyd.

Gfeller, J. D., Lynn, S. J., & Pribble, W. E. (1987). Enhancing hypnotic susceptibility: Interpersonal and rapport factors. *Journal of Personality and Social Psychology*, *52*(3), 586.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93.

Green, J. P. (2004). The five factor model of personality and hypnotizability: little variance in common. *Contemporary Hypnosis*, *21*(4), 161-168.

Green, J. P., & Lynn, S. J. (2010). Hypnotic responsiveness: Expectancy, attitudes, fantasy proneness, absorption, and gender. *International Journal of Clinical and Experimental Hypnosis*, *59*(1), 103-121.

Haggard, P., Cartledge, P., Dafydd, M., & Oakley, D. A. (2004). Anomalous control: when 'free-will'is not conscious. *Consciousness and cognition*, *13*(3), 646-654.

Hilgard, E. R., & Tart, C. T. (1966). Responsiveness to suggestions following waking and imagination instructions and following induction of hypnosis. *Journal of Abnormal Psychology*, *71*(3), 196.

Hilgard, J. R. (1974). Sequelae to hypnosis. *International Journal of Clinical and Experimental Hypnosis*, *22*(4), 281-298.

Jeffreys, H. (1961). *Theory of probability*, 3rd edn. Oxford: Oxford University Press.

Kihlstrom, J. F. (1985). Hypnosis. *Annual review of psychology*, *36*(1), 385-418.

Kihlstrom, J. F. (2005). Is hypnosis an altered state of consciousness or what?. *Contemporary Hypnosis*, *22*(1), 34-38.

Kihlstrom, J. F. (2008). The domain of hypnosis, revisited. *The Oxford handbook of hypnosis: Theory, research, and practice*, 21-52.

Kirsch, I. (1997). Suggestibility or hypnosis: What do our scales really measure?. *International Journal of Clinical and Experimental Hypnosis*, *45*(3), 212-225.

Kirsch, I. (2011). The altered state issue: Dead or alive?. *International Journal of Clinical and Experimental Hypnosis*, *59*(3), 350-362.

Kirsch, I., & Braffman, W. (2001). Imaginative suggestibility and hypnotizability. *Current directions in psychological science*, *10*(2), 57-61.

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). San Diego, CA: Academic Press.

Kruschke, J. K. (2010). Doing Bayesian data analysis: A tutorial with R and BUGS. Burlington, MA: Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Levinson, D. B., Stoll, E. L., Kindy, S. D., Merry, H. L., & Davidson, R. J. (2014). A mind you can count on: validating breath counting as a behavioral measure of mindfulness. *Frontiers in psychology*, *5*.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585.

Lush, P., Caspar, E. A., Cleeremans, A., Haggard, P., Magalhães De Saldanha da Gama, P. A., & Dienes, Z. (2017). The power of suggestion: posthypnotically induced changes in the temporal binding of intentional action outcomes. *Psychological science*, *28*(5), 661-669.

Lush, P., Moga, G., McLatchie, N., & Dienes, Z. (2018). The Sussex-Waterloo Scale of Hypnotizability (SWASH): measuring capacity for altering conscious experience. *Neuroscience of Consciousness*, *2018*(1), niy006.

Lush, P., Scott, R. B., Moga, G., & Dienes, Z. (2018). *Norms for a computerized version of the SWASH*. Manuscript in preparation.

Lynn, S. J., Weekes, J. R., Matyi, C. L., & Neufeld, V. (1988). Direct versus indirect suggestions, archaic involvement, and hypnotic experience. *Journal of Abnormal Psychology*, *97*(3), 296.

Martin, J. R., & Dienes, Z. (2018). *Bayes to the rescue: Does the type of hypnotic induction matter?*. Manuscript submitted for publication.

McConkey, K., Barnier, A. J., Maccallum, F. L., & Bishop, K. (1996). A normative and structural analysis of the HGSHS: A with a large Australian sample. *Australian Journal of Clinical & Experimental Hypnosis*.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172–175.

Montgomery, G. H., Duhamel, K. N., & Redd, W. H. (2000). A meta-analysis of hypnotically induced analgesia: How effective is hypnosis?. *International Journal of Clinical and Experimental Hypnosis*, *48*(2), 138-153.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, *23*(1), 103-123.

Morgan, A. H., & Hilgard, E. R. (1973). Age differences in susceptibility to hypnosis. *International journal of clinical and experimental Hypnosis*, *21*(2), 78-85.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition*, *25*, 27–41.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, *231*(694-706), 289-337.

Oakley, D. A. (2006). Hypnosis as a tool in research: experimental psychopathology. *Contemporary Hypnosis*, *23*(1), 3-14.

Oakley, D. A., & Halligan, P. W. (2009). Hypnotic suggestion and cognitive neuroscience. *Trends in cognitive sciences*, *13*(6), 264-270.

Oakley, D. A., & Halligan, P. W. (2013). Hypnotic suggestion: opportunities for cognitive neuroscience. *Nature Reviews Neuroscience*, *14*(8), 565.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *118*, 155–164.

Page, R. A., & Green, J. P. (2007). An update on age, hypnotic suggestibility, and gender: a brief report. *American Journal of Clinical Hypnosis*, *49*(4), 283-287.

Palfi, B., Parris, B. A., McLatchie, N., Kekecs, Z., & Dienes, Z. (2018). *Can unconscious intentions be more effective than conscious intentions? Test of the role of metacognition in hypnotic response.* Cortex, (Stage 1 Registered Report)

Parris, B. A., Dienes, Z., Bate, S., & Gothard, S. (2014). Oxytocin impedes the effect of the word blindness post-hypnotic suggestion on Stroop task performance. *Social cognitive and affective neuroscience*, *9*(7), 895-899.

Patton, J. H., & Stanford, M. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774.

Perry, C., & Laurence, J. R. (1980). Hypnotic depth and hypnotic susceptibility: A replicated finding. *International Journal of Clinical and Experimental Hypnosis*, *28*(3), 272-280.

Piccione, C., Hilgard, E. R., & Zimbardo, P. G. (1989). On the degree of stability of measured hypnotizability over a 25-year period. *Journal of Personality and Social Psychology*, *56*(2), 289.

Polito, V., Barnier, A. J., & Woody, E. Z. (2013). Developing the Sense of Agency Rating Scale (SOARS): An empirical measure of agency disruption in hypnosis. *Consciousness and cognition*, *22*(3), 684-696.

Raz, A. (2011). Hypnosis: a twilight zone of the top-down variety: Few have never heard of hypnosis but most know little about the potential of this mind–body regulation technique for advancing science. *Trends in cognitive sciences*, *15*(12), 555-557.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). San Diego, CA: Academic Press.

Raine, A., & Benishay, D. (1995). The SPQ-B: a brief screening instrument for schizotypal personality disorder. *Journal of personality disorders*, *9*(4), 346-355.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195-223.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225-237.

Rudski, J. M., Marra, L. C., & Graham, K. R. (2004). Sex differences on the HGSHS: A. *International Journal of Clinical and Experimental Hypnosis*, *52*(1), 39-46.

Shor, R. E., & Orne, E. C. (1963). Norms on the Harvard Group Scale of Hypnotic Susceptibility, Form A. *International Journal of Clinical and Experimental Hypnosis*, *11*(*1*), 39‑47.

Spanos, N. P., Radtke, H. L., Hodgins, D. C., Stam, H. J., & Bertrand, L. D. (1983). The Carleton University Responsiveness to Suggestion Scale: normative data and psychometric properties. *Psychological Reports*, *53*(2), 523-535.

Terhune, D. B., & Cardeña, E. (2016). Nuances and uncertainties regarding hypnotic inductions: toward a theoretically informed praxis. *American Journal of Clinical Hypnosis*, *59*(2), 155-174.

Terhune, D. B., Cleeremans, A., Raz, A., & Lynn, S. J. (2017). Hypnosis and top-down regulation of consciousness. *Neuroscience & Biobehavioral Reviews*.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, *25*(1), 35-57.

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169-176.

Wagstaff, G. F., Cole, J. C., & Brunas-Wagstaff, J. (2008). Measuring hypnotizability: The case for self-report depth scales and normative data for the Long Stanford Scale. *Intl. Journal of Clinical and Experimental Hypnosis*, *56*(2), 119-142.

Weitzenhoffer, A. M., & Hilgard, E. R. (1962). *Stanford hypnotic susceptibility scale, form C* (Vol. 27). Palo Alto, CA: Consulting Psychologists Press.

Weitzenhoffer, A. M. (1974). When is an "instruction" an "instruction"?. *International Journal of Clinical and Experimental Hypnosis*, *22*(3), 258-269.

Weitzenhoffer, A. M. (1980). Hypnotic susceptibility revisited. *American Journal of Clinical Hypnosis*, *22*(3), 130-146.

Wilson, S. C., & Barber, T. X. (1978). The Creative Imagination Scale as a measure of hypnotic responsiveness: Applications to experimental and clinical hypnosis. *American Journal of Clinical Hypnosis*, *20*(4), 235-249.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological methods*, *12*(4), 399-413.